TRANSLATIONAL BIOINFORMATICS FOR PERSONALIZED MEDICINE AND INTEGRATIVE BIOLOGY: DATA INTEGRATION, EXTRACTION, KNOWLEDGE DISCOVERY AND VISUALIZATION

A Thesis
Presented to
The Academic Faculty

by

Karan Uppal

In Partial Fulfillment
of the Requirements for the Degree
PhD in the
School of Biology

Georgia Institute of Technology
August 2015

TRANSLATIONAL BIOINFORMATICS FOR PERSONALIZED MEDICINE AND INTEGRATIVE BIOLOGY: DATA INTEGRATION, EXTRACTION, KNOWLEDGE DISCOVERY AND VISUALIZATION

Approved by:

Dr. Eva K. Lee, Advisor
School of Indusrial and Systems Engineering
Georgia Institute of Technology

Dr. Yuhong Fan
School of Biology
Georgia Institute of Technology

Dr. Shamkant B. Navathe
School of Computer Science
Georgia Institute of Technology

Dr. Dean P. Jones
Department of Medicine
Emory University

Dr. King Jordan
School of Biology
Georgia Institute of Technology

Date Approved: July 20, 2015

This thesis is dedicated to my parents, Anil K. Uppal and Versha Uppal.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Eva K. Lee, for being a great mentor and helping me discover my strengths by giving me the opportunity to work on a variety of projects and allowing me to participate in research efforts like Center for Health Organization and Transformation. I want to thank Dr. Shamkant Navathe, Dr. King Jordan, and Dr. Yuhong Fan for serving on my committee, giving me useful feedback, and helping me improve my research.

I want to thank Dr. Dean P. Jones (Emory University) for his tremendous support and guidance over the last four years.

I wish to thank my parents (Anil K. Uppal and Versha Uppal), my brother (Konark Uppal) and sister-in-law (Gurminder S. Uppal), my girlfriend (Samantha Latty), my nephew (Kuber Uppal) and niece (Kiara Uppal), and my close friends (Karan Budhiraja and Chad Willems) for always being there for me and supporting me. I would also like to the rest of my family (Mr. KK Rawal, Santosh Rawal, Dr. Ashok Behl, and all others) for their good wishes and support. I would like to thank my late grandparents (paternal and maternal) for their blessings and guidance.

Lastly, I want to thank past and current members of the Clinical Biomarkers laboratory at Emory University (Young-Mi Kang, Youngja Park, Shuzhao Li, James Roede, ViLinh Tran, Douglas Walker, Michael Orr, Bill Liang, and all others) for their help and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

SUMMARY

This thesis focuses on developing a computational framework to support the Precision Medicine Initiative. The newly developed tools and algorithms use machine learning, text mining and visualization techniques for extracting salient information from heterogeneous sources such as scientific literature, clinical text, and –omics technologies to enhance clinical decision making and improve the quality of healthcare. Various advances in biomedical technologies have enhanced our ability to study disease processes at different molecular levels (genes, metabolites, histones, etc.). Similarly, technological advances in healthcare domain such as adoption of Electronic Health Record systems (EHRs) provide us a unique opportunity to develop a learning healthcare system where intelligent tools and algorithms can be utilized to extract information from clinical notes, patient medication records, laboratory results, etc. for early detection of medical risks and prevention of adverse drug events. The key novel contributions of this thesis are: a) development of novel full-text summarization algorithms that have been incorporated into a web application (CoReViz) for visualizing clinically relevant information and extracting relevant sentences from clinical text and scientific articles; b) development of novel association mining algorithms and graph summarization techniques incorporated into a web application (SEACOIN2.0) for interactive drill-down summarization and hypothesis generation to extend the functionality of PubMed; c) introduction of the concept of literature based Phenotype-wide Association Studies (Lit-PheWAS); d) development of an ensemble feature selection framework for biomarker discovery using agent-based modeling and stochastic optimization techniques.

CHAPTER 1

INTRODUCTION

In January 2015, President Obama launched the Precision Medicine Initiative that aims to develop treatment strategies (initially for cancer and diabetes) taking into account individual variability (www.whitehouse.gov/precisionmedicine). The initiative is still in the planning stages and it is anticipated that new technologies and research efforts will be needed to transition this from the planning stage to implementation stage. Various conceptual models of personalized medicine have been presented that discuss the potential of electronic ancillary systems for linking the patient's molecular profile (SNPs, metabolites, gene expression, etc.) with their accumulated medical information in the Electronic Health Records (EHRs) (Starren 2013).

The national EHR mandate provides financial benefits to the institutions adopting the EHRs for meaningful use to improve the quality of healthcare. According to a recent study, over 1 billion patient visits are expected to be document in EHRs in the US (Hripcsak 2013, JAMIA). Vast amount of information creates opportunities for:

- EHR-driven research to learn from patient's history and enhance clinical decision making
- mining of patients' timeline to look for relationships between clinical events (complications, treatments, admission, follow-up, before-surgery, after-surgery)

Similarly, the amount of information in biomedical databases is growing at an exponential rate due to various advances in biomedical research as a result of high-throughput –omics technologies (Jensen 2006). Over 25 million articles are available in

PubMed today. The ocean of information in both clinical and biomedical databases creates a unique opportunity for combining existing knowledge in literature with patients' molecular level information, clinical data, family history, and medical history to make intelligent decisions and improve the quality of healthcare. Although an achievable task, the information challenges due to vast amount of information in clinical and biomedical databases need to be addressed in order to make meaningful use of this ocean of information (Feblowitz 2011, Davidoff 2011). The main aim of Translational Bioinformatics is to enhance the ability to analyze and interpret large volumes of clinical and biomedical data to for preventive and predictive medicine (https://www.amia.org/applications-informatics/translational-bioinformatics).

This thesis focuses on the development of novel text mining and machine learning tools/algorithms for automated summarization of clinical and biomedical text, literature based association studies and hypothesis generation, and identification of discriminatory variables (genes, metabolites, etc.) from -omics technologies/clinical measurements for predictive modeling.

1.1 Background

1.1.1 Electronic Health Records

Electronic Health records are defined as longitudinal repository of a patients' medical information including medications, medical history, family history, imaging, clinical narratives such discharge summaries and admission notes (Jensen 2012). Recent studies have highlighted the application and importance of EHRs for EHR-driven genomics research, improving patient phenotyping, pharmacovigilance, pharmacogenomics, and improving the quality of clinical discharge notes (Jensen 2012, Denny 2013, Cohen

2013). According to the 2003 Institute of Medicine report, the main core functional areas of EHRs are:

1) Health information and data
2) Patient support
3) Results management
4) Electronic communication and connectivity
5) Decision-support management
6) Reporting and population health
7) Computerized order entry/management
8) Administrative processes

However, recent publications have expressed concerns regarding current EHR documentation policies and practices being more focused on insurance and legal requirements rather than being patient-centric (Cusack 2013, AMIA 2011 policy meeting).

1.1.2 Text mining: information retrieval, extraction, summarization

Text mining is an established field that aims to extract information and patterns from structure and unstructured text by means of statistical techniques (Cohen 2013). The structure text includes scientific literature and domain-specific text such as discharge summaries, while the unstructured text includes free-text such clinical notes, etc. This is different from natural language processing where the goal is to extract the lexical meaning of terms with respect to their syntactic (grammatical and language rules) and semantic (meaning) associations.

There are seven main sub-categories of text mining:

a) Preprocessing: This is generally the first step in the text mining workflow and involves text preprocessing steps such as stop words removal (removing commonly occurring words), stemming (

b) Information retrieval: This is generally the first step in the text mining workflow and involves retrieval of most relevant articles, documents, or web sites related to a user's query. PubMed and Google search engine are some popular information retrieval systems. PubMed is one of the largest publically available repositories of scientific literature. The system allows users to retrieve literature based on the ranking criteria such as most recent publications and recently introduced "Sort By Relevance" option that weighs search results based on overlap with query terms, citation counts, and gives higher weight to recent publications. The Entrez e-utilities provide a programming interface for querying PubMed and retrieving relevant articles using programming languages such as Perl and Java.

c) Information extraction: The next step in the text mining workflow is the process of named entity recognition to extract entities of interest such as genes, chemicals, diseases, mutations, etc. in the retrieved documents. Over the years, several text-mining challenges have focused on the problem of named entity recognition in the biomedical domain. The PubTator service uses the tmChem, tmVar, DNorm, MeSH, and OMIM tools for named entity recognition (Wei 2013). Each PubMed entry is tagged with occurrences of bioconcepts such as genes, chemicals, diseases, mutations, and species. The RESTful API provides a programming interface for querying PubTator. PubTator annotations are being used as benchmakrs in the ongoing BioCreative V challenge. Most tools use UMLS

concepts for named entity recognition; however the UMLS has limited coverage of specialized terminology such as genetic variants, which makes them less appealing for EMR driven genomics analyses or Phenotype-wide Association Studies (Plaza 2013, Denny 2013). A recent evaluation of the performance of UMLS concepts in clinical narratives showed that most clinical terms are either not represented or poorly represented in UMLS (Freidlin 2011).

d) Association mining: Once the relevant terms/concepts/entities have been identified, the next step involves relation extraction based on occurrences of biomedical entities or concepts. The criteria for association rule could be statistical or semantic relationships determined using tools like MetaMap that are based on UMLS (Aronson 2001, Liu 2006). Several association measures have been developed over the years including but not limited to tf-idf, pointwise mutual information, odds ratio, etc. (Wren 2004, Esteban 2009).

e) Summarization: Automated text summarization is the process of condensing the original text into a shorter form by extracting most salient information. The summarization process can be categorized as extractive vs abstractive and indicative vs informative (Pivovarov 2015). Extractive summarization uses a sentence ranking criteria to identify salient sentences from the original text, while abstractive summarization uses natural language processing techniques to synthesize new text based on the content of the original document. Indicative summarization provides an overview of the original document by highlighting

relevant terms/concepts, while informative summarization replaces the original text by generating summaries that cover the content in the original text.

f) Hypothesis generation/knowledge discovery: The concept of literature-based discovery was pioneered by Swanson in the 1980s when he discovered the association between fish oil and Raynaud's disease from disjoint literature using blood viscosity as the common link (Swanson 1986). This is termed as the A-B-C model, where AB and BC are disjoint sets of literature. A novel connection between A and C is made based on the common link B. Various tools have been developed since then such as Arrowsmith (Smalheiser 2009) for hypothesis generation.

Various tools have been developed over the last few decades for mining both clinical and biomedical text. Most tools for biomedical literature focus on specific relations, e.g. gene-gene, gene-disease, etc. However, the ability to summarize the large body of scientific literature and obtain a systems level overview of the interactions without running into the problem of "information overload" still remains a challenge.

1.2.2 Feature selection techniques and biomarker discovery

Feature selection is a critical step in the biomarker discovery process. Feature selection methods can be classified as: filter, wrapper, and embedded (Figure 1.1; Saeys 2007). The filter methods use statistical criteria independent of the learning machine to select relevant features and these features are then used to build/train the model. Methods such as t-test, ANOVA, F-test, Chi-sq test, mutual information, etc. can be classified as filter methods. The wrapper methods use a search strategy to evaluate different combinations

of subsets of features and select the best model based on the evaluation using a learning algorithm such as Support Vector Machine (SVM, Vapnik 1998). Different search algorithms such as best subset, genetic algorithms, PSO, etc. can be used for finding the optimal set of features; however these methods are prone to over-fitting (Saeys 2007, Christin 2010).

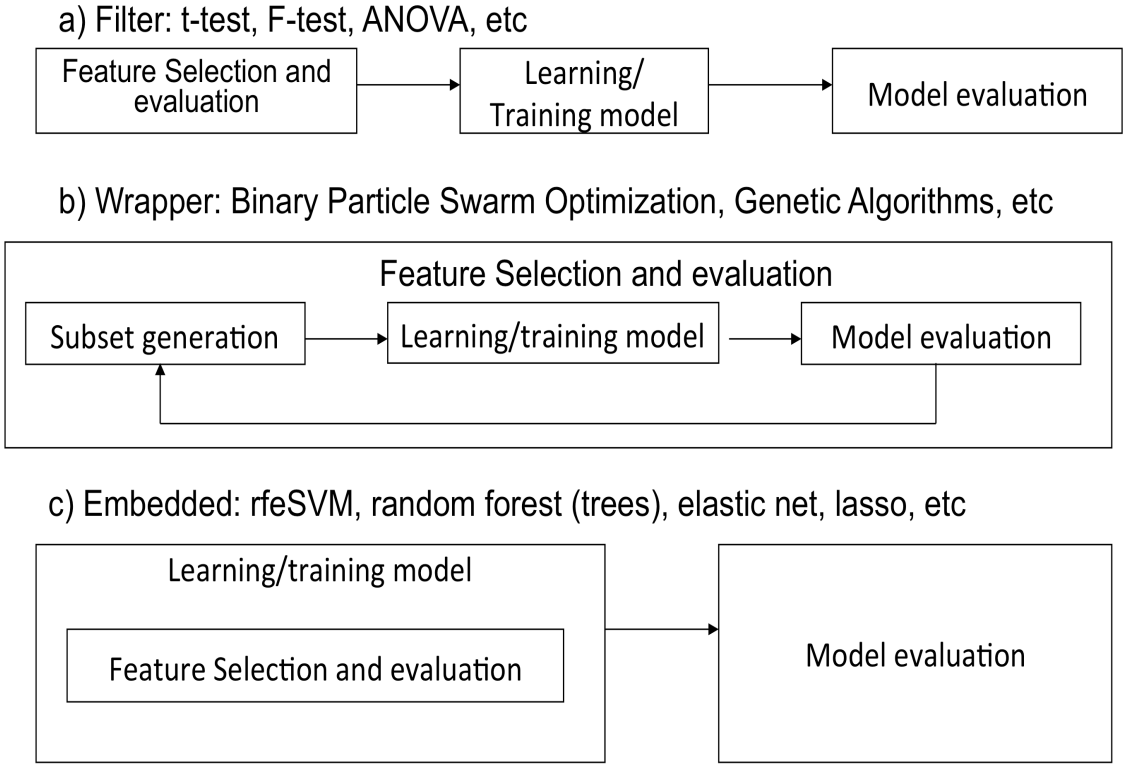

Figure 1.1. Different categories of feature selection methods

The embedded methods such as recursive feature elimination based on SVM, random forests (RF), LASSO built-in variable selection. For instance, LASSO is a coefficient shrinkage method and uses a L1 penalty function to assign a value greater than 0 if a feature is relevant, and 0 otherwise (Saeys 2007).

1.2.3 Supervised and unsupervised learning

Data can exist in either labeled or unlabeled form. In the labeled form, each instance (sample, subject, patient, etc.) is assigned to a particular class, e.g. “Cancer”, “Normal”, “Obese”, etc. These labels are used by learning machines (classifiers) for building predictive models based on the feature matrix (vector of variables, genes, metabolites, etc. per sample) that can be used for predicting the class of future or unseen samples. This is called supervised learning where the training data is available and a classifier (Support Vector Machine, Random Forests, etc.) is used for learning the classification model. Unsupervised techniques are used when the samples/instances are unlabeled. Various clustering techniques such as hierarchical clustering analysis, k-means clustering, etc. can be used for clustering the samples based on the feature matrix.

1.2.4 Thesis overview

The three novel text-mining algorithms and software systems developed as part of this thesis are introduced in Chapters 2,3, and 4. Chapter 5 gives an overview of the applications of the newly developed methods and how the newly developed tools can improve basic and translational research and enhance clinical decision-making to improve the quality of healthcare.

The remaining chapters of this thesis cover the following:

1) Chapter 2: Development of text mining algorithms and tools for automated summarization of full-text scientific articles and clinical free-text such as operative notes, radiology reports, discharge summaries, etc.
2) Chapter 3: Development of a text mining algorithms and visualization tools to extend the functionality of PubMed for performing automated summarization,

Association mining and hypothesis generation for enhancing information access to facilitate improved experimental designs, knowledge discovery, enhancing biomedical decision-making, and improved understanding of biological mechanisms at the systems level (gene, chemical, disease, mutations, etc.).

3) Chapter 4: Development of novel feature selection framework for various applications such as identifying biomarkers in omics studies and identifying discriminatory terms/words for document classification
4) Chapter 5: Applications
5) Chapter 6: Conclusion and Future Work

CHAPTER 2

COREVIZ: AN INTERACTIVE CONTENT RECOGNITION AND VISUALIZATION TOOL FOR CLINICAL AND BIOMEDICAL TEXT

Abstract

Objective: Automated summarization of scientific literature and patient records is essential for enhancing clinical decision-making and facilitating precision medicine. Most existing automated summarization methods are based on single indicators of relevance, offer limited capabilities for information visualization, and do not account for user specific interests. The aim of this research is to develop an interactive content recognition system that combines machine learning and visualization techniques with domain knowledge for highlighting and extracting salient information from clinical and biomedical text.

Methods: Extractive summarization is performed using MINTS, a novel sentence-ranking framework presented in this work that uses a random forests classifier and multiple indicators of importance for relevance evaluation and ranking of sentences. Indicative summarization is performed using the weighted TF-IDF scores of over-represented domain-specific terms. The performance of the MINTS algorithm is evaluated on 32 full-text Medline articles and 30 clinical case reports using the ROUGE toolkit.

Results: The random forests model classified sentences as “good” or “bad” with an 87.5% accuracy on the test set. Summarization results from the MINTS algorithm achieved higher ROUGE-1, ROUGE-2, and ROUGE-SU4 scores as compared to

methods based on single indicators such as term frequency distribution, position, and eigenvector centrality (LexRank) as well as random selection, $p<0.01$for both corpuses.

Conclusion:

We have developed a web-based summarization and visualization tool, CoReViz (https://newmedicalor.isye.gatech.edu/content-recognition/), for extracting salient information from clinical and biomedical text. The system ranks sentences by relevance and includes features that can facilitate early detection of medical risks in a clinical setting. The interactive interface allows users to filter content and edit/save summaries. The evaluation results on two test corpuses show that the newly developed MINTS algorithm outperforms methods based on single characteristics of importance.

2.1 Introduction

Implementation of electronic health record systems (EHRs) across healthcare institutions and growing information in biomedical databases provides a unique opportunity to enhance clinical decision-making by linking key details from patient records and clinically relevant information from the literature into the clinical decision-making workflow (Davidoff 2011). However, this is a challenging task due to the exponential growth in the amount of information in both clinical and biomedical databases. Intelligent informatics tools and algorithms are needed to automate processing of large amounts of text and address the problem of "information overload" (Bawden 2008, Smith 2010).

According to a recent review, almost half of the questions related to patient care raised by clinicians are not pursued due to limited amount of time at point of care and doubts about availability of information (Cohen 2005). Although most scientific articles

include abstracts, recent studies have shown the advantages of using full-text for summarization as not all relevant information can be reported in the abstracts (Plaza 2013). Moreover, different readers may find different pieces of information in the text useful (Del Fiol 2014).

The problem of information overload is also associated with EHRs since the amount of clinical information per patient could be excessive, particularly for patients suffering from chronic illness and multi-morbidities (Bawden 2008, Reichert 2010, Duftschmid 2013). A cognitive study of the thought process of eight physicians during the EHR review process showed that the majority of the time is spent reviewing the "Notes" section to identify problems, medical history, medications, etc. (Reichert 2010). Various studies have shown the potential of using text mining and natural language processing techniques for enhancing clinical-decision making and improving the quality of healthcare (Wilson 1998, Harvey 2003, Wang 2009). For instance, studies have shown the utility of text mining and natural language processing techniques to facilitate detection of adverse drug events and comorbidities in EHRs (Wang 2009, Salmasian 2013). It has also been shown that high-information clinical findings appear in the medical records of the patients before the high-risk diagnosis is determined (Feldman 2012). Furthermore, automated summarization of patient information to extract salient information can improve decision-making and reduce the risk of information overload (Feblowitz 2011, Pivovarov 2015).

In this work, we focus on the development of machine learning based automated text summarization techniques to address the challenges of "salient detection" and "information overload" in the healthcare and biomedical domain (Bawden 2008,

Pivovarov 2015. Automated summarization is the process of extracting important information from the original text and presenting it in a condensed form (Das 2007, Nenkova 2012, Pivovarov 2015). Summarization methods can be classified as extractive versus abstractive (Mani 1999, Pivovarov 2015). Extractive summarization involves extracting important sentences from the input text according to a scoring or ranking criteria, while abstractive methods use natural language processing techniques to construct new sentences (Mani 1999, Pivovarov 2015). The two categories can be further classified as indicative versus informative where the indicate summaries only provide an overview of the underlying information, while informative summaries provide enough details to replace the original text (Pivovarov 2015). Various methods for extractive summarization have been developed over the last decade (Das 2007, Nenkov 2012). These methods utilize a variety of sentence ranking strategies such as intermediate topic representation, graph-based methods based on Google's PageRank algorithm and UMLS semantic relations in UMLS (http://www.nlm.nih.gov/research/umls/), MeSH terms, sentence position, and semantic relations of biomedical concepts (Herskovic 2011). For example, Bhattacharya et al. demonstrated that usage of MeSH terms improves summarization results, Fiszman et al. used semantic relationships for summarization of Medline citations, Reeve et al. used the concept frequency for summarization, Jonnalagadda et al. use UMLS concepts and TextRank algorithm for extracting sentences related to a particular topic from Medline abstracts, and Mishra et al. clinically relevant sentences from UpToDate (Reeve 2006, Das 2007, Fiszman 2008, Bhattacharya 2011, Herskovic 2011, Nenkova 2012, Jimeno-Yepes 2013, Mishra 2013). Most existing extractive summarization methods utilize single indicators of relevance for sentence

ranking that might not be relevant for all types of clinical and biomedical. The data mining and exploration process can be made more effective by incorporating human knowledge and allowing the user to interact with the system using visualization methods that provide an insight into the underlying information (Keim 2002, Feblowitz 2011, Hirsch 2015).

Here we present CoReViz, a content recognition and visualization tool that uses a multi-stage sentence evaluation and ranking framework for extracting salient information from the input text. A random forests classifier is used in stage one for evaluating worthiness ("important" versus "not important" for summarization) of each sentence in the input text. In the next stage, a rank aggregation scheme based on multiple indicators is used for identifying the best set of sentences to be included in the final summary. The performance of the multi-stage summarization scheme was evaluated against existing summarization techniques using a subset of articles from the CRAFT corpus and a corpus of full-text clinical case reports obtained from Medline (Bada 2012). Indicative summarization is performed using an interactive topic cloud based on the over-represented biomedical terms in the input text. The topic cloud provides a visual overview of the content in the input text and allows interactive filtering of the sentence extraction results based on users' interests. A keyword based filtering allows users to generate a summary based on the top-ranked sentences and edit and save the selected summary for future review or additional processing such as language translation. Finally, related articles in PubMed are presented based on the topic cloud to incorporate external knowledge.

The main objectives of this research are: 1) development of extractive and indicative summarization algorithms to address information challenges related to precision medicine; 2) development of a web-based interactive summarization tool that accounts for user specific interests and can facilitate clinicians in summarizing clinical/biomedical text by highlighting key information both at the level of individual terms and sentences.

2.2 SYSTEM AND METHODS

The system is developed using PHP, ADOBE Flex, Java, MySQL, and Apache Lucene. An overview of the system and details of each processing step are provided below. Figure 2.1 gives an overview of the CoReViz system.

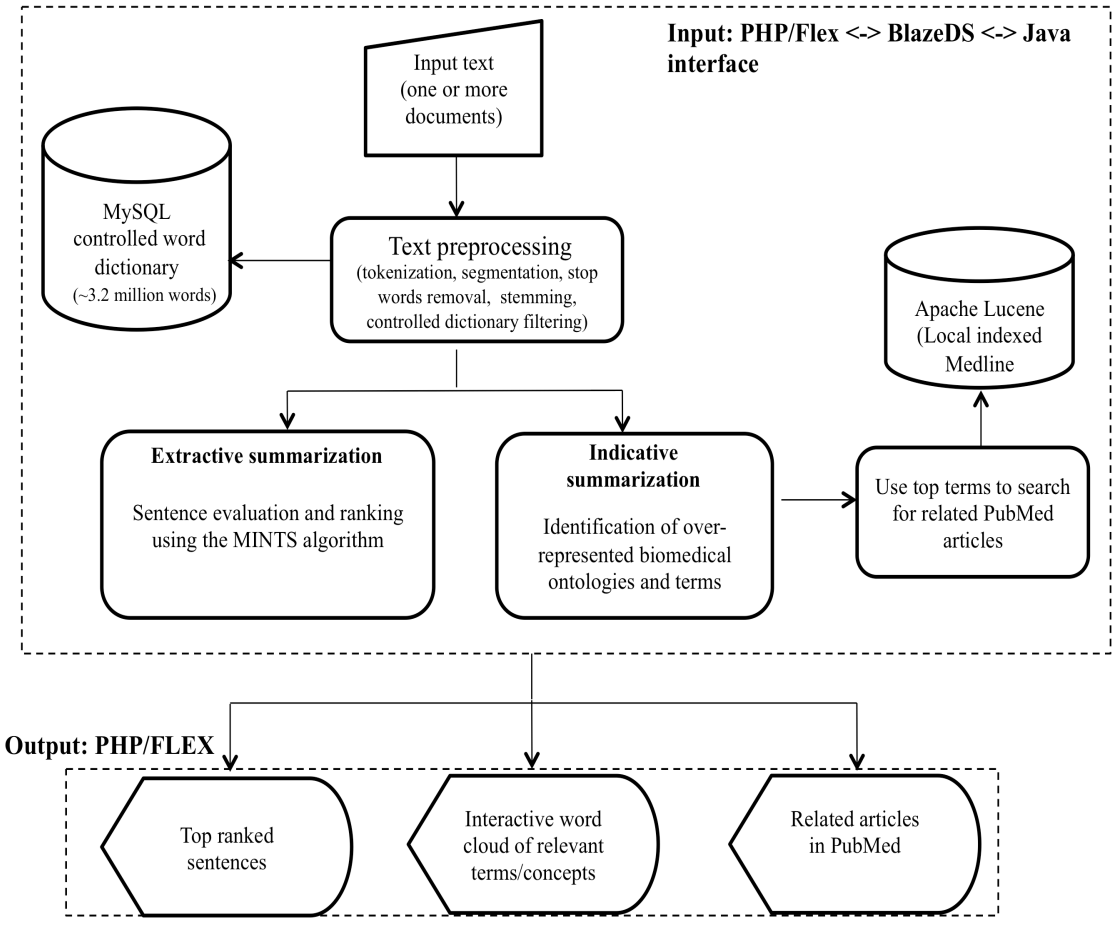

Figure. 2.1. CoReViz system diagram

2.2.1 Preprocessing: Segmentation, tokenization, and stemming

Segmentation of input text into individual sentences is performed using the LingPipe tool kit (http://alias-i.com/lingpipe/). Regular expression rules are used to segment the input text into word tokens. Porter stemmer algorithm is used to reduce all inflected forms of a word to the same text string, eg: {densities, density} -> density (Porter 1980).

2.2.2 Indexed database of Medline abstracts

An indexed database of Medline abstracts published between 1975 and 2015 was generated using Apache Lucene (http://lucene.apache.org). Lucene is a Java-based text search engine that facilitates efficient querying and document retrieval.

2.2.3 Dictionary of controlled vocabulary and stop words

A controlled dictionary of 3.2 million words was generated using MeSH terms, SNOMED-CT, and PubTator, which includes terms related to genes, proteins, genetic variants, taxonomy, diseases/disorder, and chemicals from biomedical literature (http://www.nlm.nih.gov/research/umls/Snomed , Rogers 2012, Wei 2013). In addition to the 121 stop words used by PubMed (http://www.ncbi.nlm.nih.gov/pubmed), any word in the input text that is not present in the controlled dictionary is considered a stop word.

2.2.4 MINTS: a multi-stage algorithm for sentence extraction and ranking

Various machine learning (ML) techniques such as Naïve Bayes classifier, decision trees, and hidden Markov Model (HMM), etc. have been implemented for sentence extraction (Das 2007, Nenkova 2012). Some of the common importance indicators utilized by previous ML methods include sentence length, position, TF-IDF, and parts of speech (Das 2007). In this work, we have developed a three-stage procedure to extract relevant sentences using a random forests (RF) classifier and

various indicators of relevance such as: sentence length, position in the input text, number of clinical/biomedical terms, percentage of terms that are clinical/biomedical terms, normalized degree centrality, and overlap with global term frequency distribution determined using the Dice-coefficient as similarity metric (Breiman 2001).

In stage one, a sentence-feature matrix is generated where each row corresponds to an individual sentence and the columns represent the indicators of relevance. The number of domain-specific terms is determined using the controlled dictionary. A TF-IDF based cosine similarity matrix is used to determine the degree centrality of each sentence, which is normalized by the total number of sentences in the input text.[18] According to Luhn's theory, the most frequent terms/concepts are the most important ones and can be used to determine the significance of individual sentences.[19,35] The overlap between the term frequency distribution of the current sentence and the global frequency distribution is determined using Dice-coefficient (1) as similarity metric, which has been previously shown to outperform other similarity functions metrics for determining the overlap between a candidate summary and source text (Reeve 2006).

$$\text{Score}\,(s) = 2 * \frac{|A \cap B|}{|A| + |B|} \qquad (1)$$

where

s = index of current sentence,

$|A|$ = number of relevant terms/concepts in the frequency distribution model of the entire document,

$|B|$ = number of relevant terms/concepts in the frequency distribution model of sentence s,

$|A \cap B|$ = number of overlapping terms/concepts between the global frequency distribution model and the distribution model of sentence s

In stage two, a random forests classifier is used to predict the "worthiness" of a sentence. Random forests is a non-parametric supervised classification technique that uses an ensemble of decision trees for learning a model (Breiman 2001). Each tree in the forest is generated using a random set of variables (relevance indicators) and by sampling a random set of training samples (bagging). The trees are grown until the leaves/terminal nodes contain samples belonging to the same class. After the forest is constructed, every tree casts a vote for the class assignment of the new sample. The majority vote is used to determine the class of the new sample. The randomForest package in R is used in CoReViz.

35 full-text articles from the CRAFT corpus were used for building the RF model. The remaining 32 articles were used during the evaluation stage as described later. The two sets will be termed as CRAFTtrain and CRAFTtest, respectively. All articles were preprocessed to remove stop words. Each sentence in the training set was annotated as "good" or "bad" based on the amount of overlapping information between sentence, s_i, and the article abstract. For each sentence, s_i, a Dice-Sorenson (DS) index (1) was used to determine its overlap with the set of terms in the abstract. The maximum DS index values per article ranged from 0.046 to 0.358 in CRAFTtrain corpus with a median value of 0.01 across all sentences in the training corpus. All sentences with DS index less than 0.01 were annotated as "bad" or not important for summary. This resulted in 7,498 out of 9,779 sentences being annotated as "good" for summarization. A random forests model

was trained using 60% of sentences (5,867) and the performance of the classifier evaluated using the remaining 40% of sentences (3,912). This trained classifier is used to assess the importance of every sentence in the new text. And, only the sentences that are predicted as "good" are used in the ranking stage. This facilitates document compression/data reduction.

The last stage involves selection of "good" sentences for generating summaries based on aggregated ranking and redundancy evaluation. The scores based on m indicators for every sentence i are converted to ranks, R_{im}=[1…N], where i is the index of sentence s_i, m is the indicator of importance (e.g. degree centrality, position, etc.), and N is the number of sentences. An aggregated rank is assigned to each sentence by taking the average of rankings from different indicators. The top ranked sentences are used for summaries after evaluating the cosine similarity (a threshold of 0.4 is used based on empirical evaluation) between the previously selected sentences in the summary set and the incoming sentence to reduce redundancy. We call this new algorithm Multi Indicator Text Summarization or MINTS (Figure 2.2). A normalized score ranging between 0 (least important) to 1 (most important) is assigned to each sentence. The maximum number of sentences to be selected is a user-defined parameter.

For comparison purposes, topic-based and graph-based extraction summarization techniques were also included during the evaluation process:

i. topicDist: As described above, this method evaluates the relevance of a sentence term/concept frequency based on the overlap with the most frequent terms/concepts in the entire text (Luhn 1958, Reeve 2006, Plaza 2013).

ii. LexRank: LexRank is a graph based extractive summarization approach that uses the cosine similarity matrix to determine similarity between sentences and uses eigenvector centrality to extract relevant sentences (Erkan 2004). A network of sentences is generated where each node represents a sentence, s_i, and the edges represent the cosine similarity between s_i and s_j. The LexRank algorithm implemented in the MEAD toolkit was used for evaluation (http://www.summarization.com/mead/).

2.2.5 Indicative summarization

A word-cloud based visualization method is used to represent term/concept distribution.[28] This provides a concept-oriented summarization of the over-represented relevant terms and concepts in the input text. A weighted scoring scheme is used to prioritize terms corresponding to diseases/disorders, genes, mutations, and chemical names.

$$Score(t) = i * (\mathrm{Wc}) * tf * IDF, \qquad (2)$$

where

i = 1 if the term is found in the controlled vocabulary, 0 otherwise,

W_c = 1000 if the term is a disease/disorder, chemical, mutation, gene

1 otherwise,

tf = frequency of term t in the input text,

$$\mathrm{IDF} = 1 + \log\left(\frac{(total\ number\ of\ indexed\ Medline\ abstracts)}{(number\ of\ abstracts\ with\ term\ t)}\right)$$

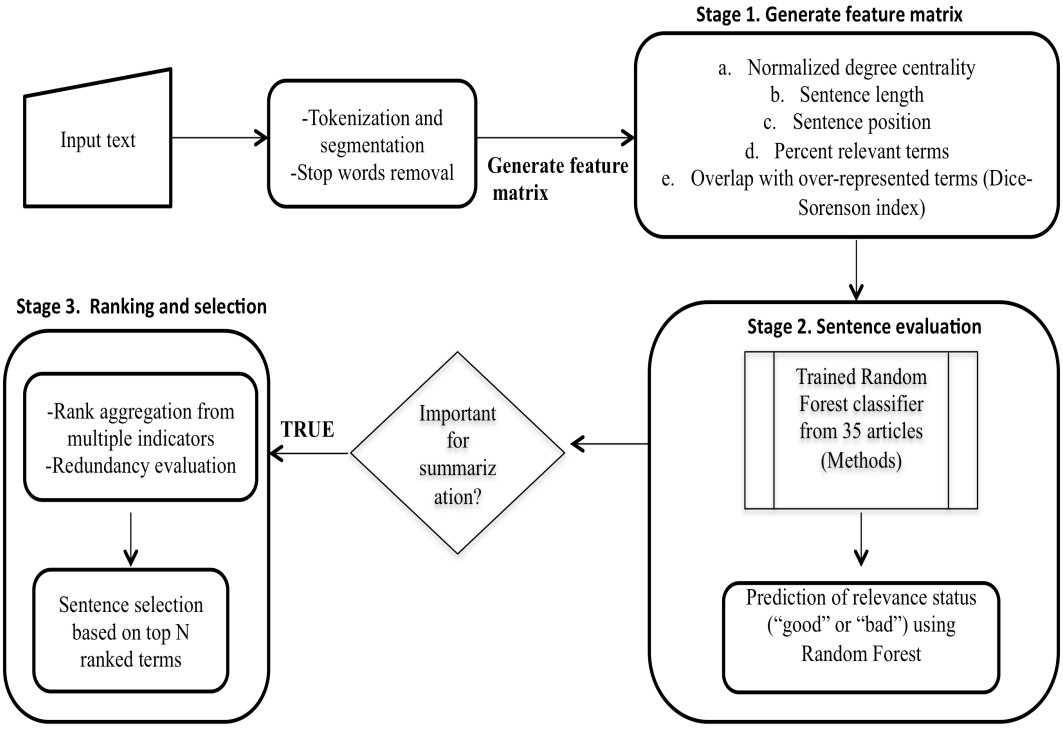

Figure 2.2. MINTS workflow for extracting salient sentences

2.2.6 Interactive or user-guided summarization

Visual data exploration provides insights into the data and makes the data mining process more effective by incorporating human perception and intelligence (Keim 2002). CoReViz facilitates visual mining by means of an interactive word cloud. The word cloud represents the distribution of the relevant terms in the input text and can be used to interactively filter the ranked list of sentences to generate keyword-based summaries. Alternatively, users can manually define the keywords for filtering the ranked sentences to generate query-specific summaries.

2.2.7 Evaluation strategy

A set of 32 full-text articles from the CRAFTtest corpus and a set of 30 randomly selected full-text clinical case reports from BMC Ophthalmology, BMC Neurology, BMC Pulmonary Medicine, BMC Cancer, and New England Journal of Medicine were used to evaluate the performance of the three sentence ranking methods: MINTS, LexRank, and topicDist. Position-based ranking and random selection were used as baseline. In the position-based selection, sentences were assigned scores according to their position in the document such that the first sentence gets the highest score and the last sentence is assigned the lowest score. An extractive summary was generated using each method from the full-text of the articles using the top five sentences. The summaries generated by the system were compared with human generated summaries (abstracts) using ROUGE, a software package for evaluation and comparison of summaries based on the n-gram co-occurrence statistics (Lin 2003). A one-sided paired Wilcoxon signed-rank test was used to evaluate the significance of differences between the ROUGE scores for randomly generated summaries and different summarization algorithms. The average performance of three randomly generated summaries was used for comparison.

2.3 RESULTS AND EVALUATION

The random forests model achieved an out-of-bag classification accuracy of 87.78% on the training set. An overall classification accuracy of 87.5% and a balanced accuracy rate (group-specific accuracy) of 79.4% (94.6% for “good” category and 64.2% for “bad” category) was achieved for the blinded test set of 3,812 sentences. The ROUGE evaluation scores of extractive summaries generated using different methods are shown

in Figure 2.3. MINTS gave the best performance in both experiments. For the CRAFTtest corpus of 32 full-text articles, MINTS gave ROUGE-1, ROUGE-2, and ROUGE-SU4 scores of 0.414, 0.136, and 0.171, respectively, with p-values from the one-sided Wilcoxon signed rank test ranging from 10^{-13} to 10^{-8} for the three scores (Figure 2.3a). MINTS outperformed the other two methods with 15% and 38% improvement in ROUGE-1 scores as compared to the topicDist and LexRank, respectively. Both topicDist and LexRank methods performed better than the baseline.

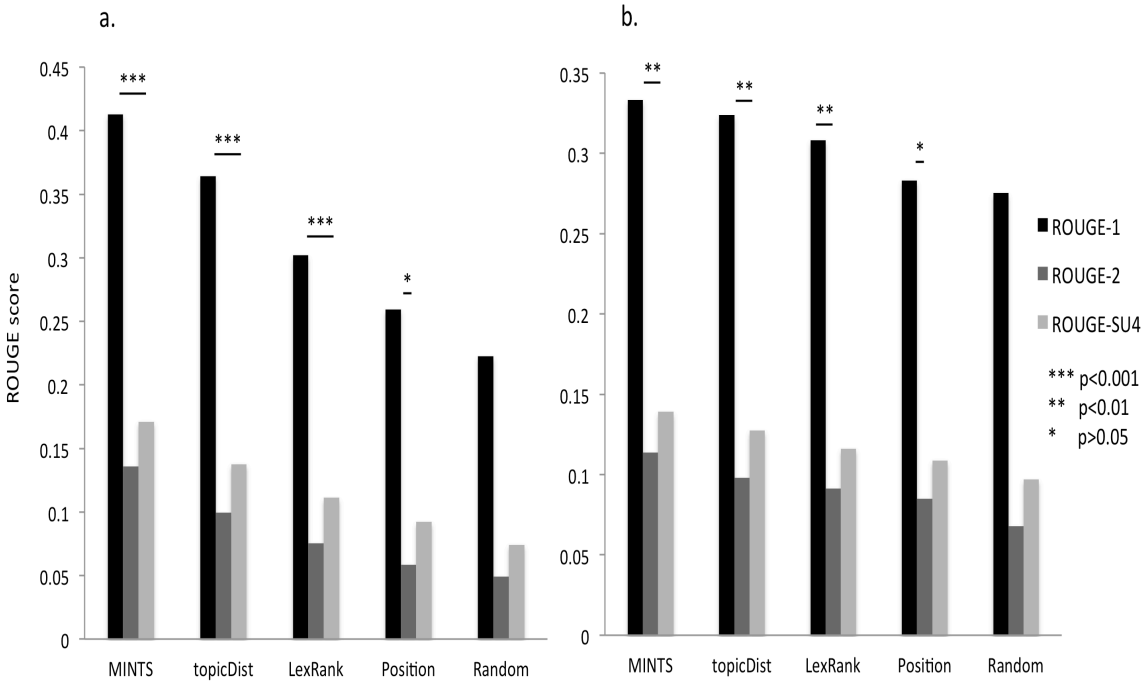

Figure 2.3. ROUGE evaluation scores for different extractive summarization methods using abstracts as gold standard summaries. a. Performance comparison of different summarization approaches on CRAFTtest corpus (32 full-text scientific articles); b. Performance comparison of different summarization approaches using ROUGE evaluation toolkit on clinical case reports corpus. The p-values for ROUGE-1, ROUGE-2, and ROUGE-SU4 scores for each method were compared against respective baseline scores using a one-sided Wilcox text. The MINTS algorithm outperformed other methods in both cases.

Similar ranking pattern was observed for the different extraction methods using the clinical case reports corpus as evaluation set (Figure 2.3b). However, the p-values were higher as compared to the CRAFTtest evaluation (0.001 to 0.01), which is likely due to the differences in the lengths of the documents in the two corpuses. The number of sentences in the clinical case reports corpus ranged from 18 to 72, while the number of sentences in the CRAFTtest corpus ranged from 101 to 455. As described in Methods, users have the option to specify the number of sentences to be extracted and used for generating the document summary.

2.4 DISCUSSION

The success of new healthcare initiatives such as the Precision Medicine Initiative relies a lot on the ability of computational tools and algorithms to address challenges related to efficient and impactful usage of information existing in different data sources. The vast amount of information in electronic health records and scientific literature has the potential to enhance clinical decision-making and improve the quality of healthcare as more informed decisions can be made at the patient level by integrating knowledge in the biomedical domain with patient characteristics and medical history (Bawden 2008, Reichert 2010, Duftschmid 2013). However, the growing sizes of biomedical and clinical databases have created the problem of "information overload" (Smith 2010). A large amount of information in the healthcare domain such as clinical notes, discharge summaries, radiology reports etc. is stored in the form of text. Most existing text summarization tools for clinical/biomedical domain utilize single indicators of relevance such as concept distribution, position, and rely upon UMLS as the main vocabulary for identifying concepts and semantic relations between concepts, which limits the

incorporation of specialized biomedical terminology such as genetic variants (Plaza 2013). In addition to natural language processing, visualization techniques are essential for representation of information in a form that facilitates pattern recognition and large volumes of data (Feblowitz 2011, Hirsch 2015).

We have developed a web-based content recognition and summarization tool, CoReViz (https://newmedicalor.isye.gatech.edu/content-recognition/), for clinical and biomedical text that includes features such as extractive summarization to identify relevant sentences, indicative summarization of the overrepresented biomedical terms and concepts in the input text using word cloud visualization, interactive concept-oriented summarization, and retrieval of biomedical literature relevant to the input text (Figure 4). A controlled vocabulary dictionary generated using MeSH, Snomed-CT, and PubTator is used for determining relevant terms.

Extractive summarization is performed using the MINTS algorithm as described in the Methods. The algorithm uses a multi-stage framework that combines supervised learning techniques, individual characteristics of sentences (position, length, relevant terms) and network level characteristics (degree centrality) for extracting salient sentences. A random forests classifier trained on a set of 9,779 sentences from 35 full-text articles from the CRAFT corpus is used for evaluating sentence worthiness for summarization, “good” vs “bad”. Multiple indicators of importance such as degree centrality, presence and number of relevant terms, and position are used during relevance evaluation and ranking stages. An aggregated ranking scheme and cosine similarity based redundancy evaluation is used for selecting top sentences. Redundancy detection is performed using cosine similarity between potential candidates and already selected

sentences. The performance evaluation results on full-text scientific articles and clinical case reports show that the summarization process can be improved by combining machine learning, text mining, network analysis techniques with domain knowledge as opposed to using single characteristics of relevance (Das 2007, Nenkova 2012). Furthermore, the results suggest that the length of the input text does not affect the performance of the MINTS algorithm. The two corpuses varied in their sizes as well as structure and content as the clinical case reports focus on diagnosis, treatment, and management of clinical cases and are targeted towards clinical audience, while the scientific articles focus on basic science or biomedical research. These results demonstrate the promise of "intelligent" algorithms like MINTS to address the problem information overload in both the clinical and biomedical domain.

The system has several additional features to enhance clinical decision-making:

- *Document-driven search to retrieve related literature from Medline*: CoReViz uses the clinically/biologically relevant terms to find related articles in PubMed. This allows users to gain additional information about the key diseases or medications that are mentioned in the input text.

- *Visualization of over-represented terms using controlled dictionary (PubTator, MESH and SNOMED CT):* The system uses the term-frequency criteria to identify clinically/biologically relevant terms in the input text. A word cloud representation of the top clinically/biologically relevant terms is generated. This could facilitate detection of high risk findings

- *Interactive interface and visualization:* The web interface allows users to generate and edit automated summaries from the ranked sentences. Users have the option to filter sentences by keywords and generate a summary of the document based on the relevant sentences.

- *Library of summaries:* The system allows the users to automatically generate, edit, and save summaries for downstream pattern mining.

Figure 4 shows an illustration of the system. Users can use the copy/paste option or upload a Word document with input text. A table with relevance scores for each sentence is returned based on the newly developed MINTS algorithm. Users can filter the sentences based on keywords, e.g. "diabetes". Alternatively, the interactive word cloud can be used for filtering the sentences by clicking on the term of interest. An extractive summary can be generated using top N sentences, where N is a user-defined parameter.

Users have the option to edit and save the generated summary for future analysis such as temporal tracking of clinically relevant indicators or medication usage. The system also provides a list of related PubMed articles based on the top over-represented terms in the input text.

Limitations: First, although the evaluation was performed on different types of full-text articles from both biomedical and clinical domain, further validation is required including extrinsic assessment by clinicians. Second, the terms in the topic cloud are currently not mapped to their corresponding concepts leading to ambiguity and redundancy if a concept is represented in different forms in the input text. Third, the random forests classifier was built using only a subset of all possible indicators of relevance leaving room for improvement at the initial sentence evaluation level.

Furthermore, the classifier was built using a training set of 35 full-text articles. Evaluation of performance of different classification algorithms using a larger training set can lead to better sentence evaluation and summarization results. Fourth, the algorithms used for indicative and extractive summarization do not utilize lexical or semantic relationships between terms/concepts. A more detailed natural language analysis could further improve the performance of the summarization algorithms. Finally, the system currently supports only English language.

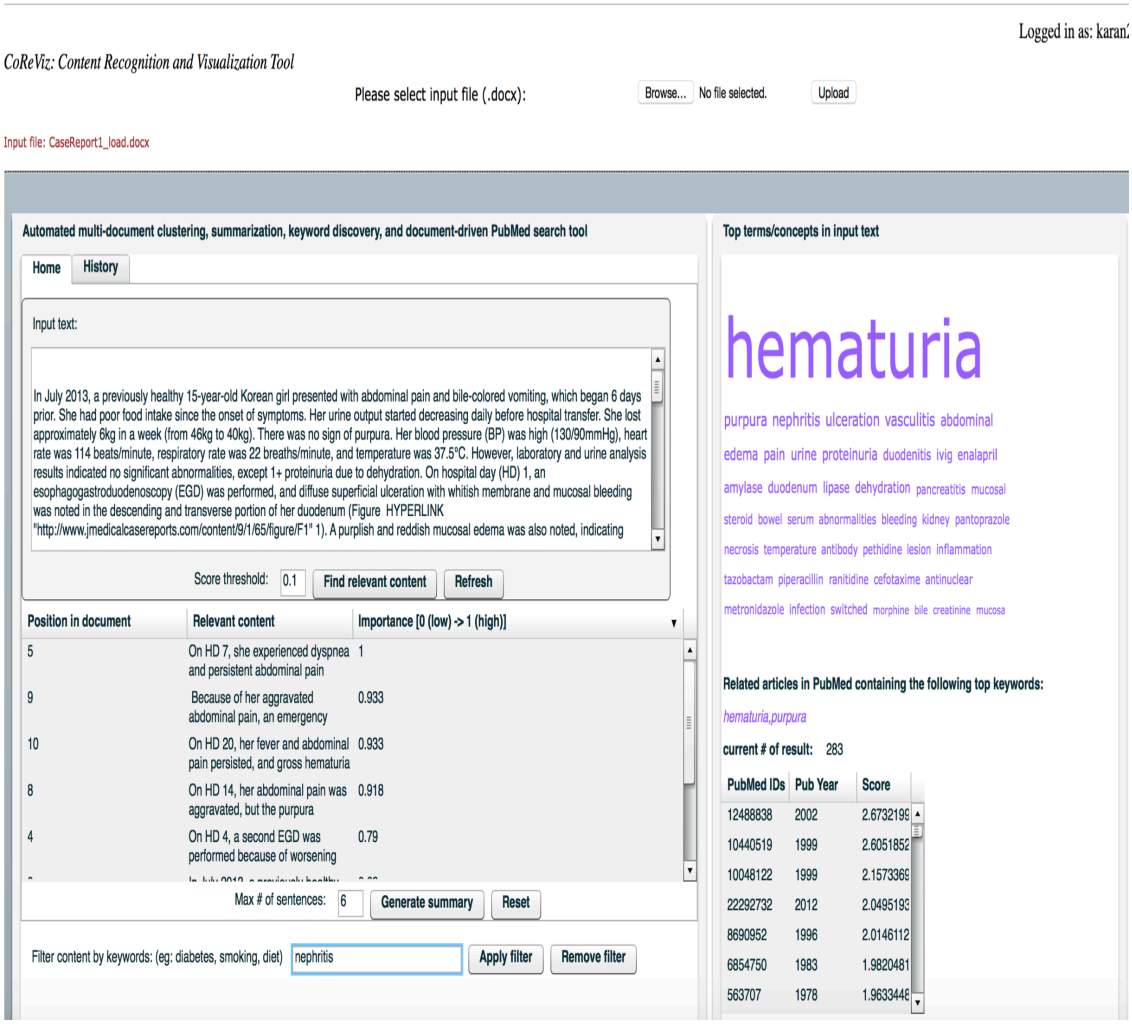

Figure. 2.4. CoReViz demonstration. Users can upload or paste the input document and select t he clustering and summarization options. The output includes a word cloud of over-represented clinical/biomedical terms, ranked sentences within each cluster, and related articles in Medline. Users can filter the list of ranked sentences based on keywords.

Future work: Extrinsic evaluation of the system and further validation of the summarization strategies using different types of clinical text such as operative notes and radiology reports will be performed in a patient care setting. The evaluation will focus on the ability of the system for high-risk findings in patient records and the impact on patient care and decision-making. The functionality of the system will be further extended by providing automated graph-based summarization of the input text as demonstrated in our previous work, SEACOIN, which was designed for topic-based summarization of

Medline abstracts (Lee 2011). The terms in the interactive cloud will be mapped to concepts in PubTator and SNOMED-CT (Wei 2013, http://www.nlm.nih.gov/research/umls/Snomed/).

2.5 CONCLUSION

Intelligent tools and techniques are required to extract information from rapidly growing data in healthcare and biomedical domain for facilitating precision medicine. In this work, we have developed CoReViz (https://newmedicalor.isye.gatech.edu/content-recognition/), an interactive content recognition and summarization tool for extracting salient information from clinical and biomedical text. The system includes both indicative and informative summarization strategies that allow the users to retrieve and visualize important content in the input text in an interactive manner. A novel multi-stage procedure, MINTS, is presented here. The algorithm uses a random forests classifier to evaluate the "worthiness" of individual sentences for summarization prior to scoring based on multiple domain specific, sentence-level, and network-level characteristics. The ROUGE evaluation results on two independent test corpuses show that MINTS provided better summarization results as compared to methods based on single indicators (topic/concept frequency distribution and LexRank). ROUGE evaluation scores for the MINTS algorithm were significantly different as compared to random selection at a significance level of 0.01: ROUGE-1 (0.41 vs 0.22), ROUGE-2 (0.14 vs 0.06), and ROUGE-SU4 (0.17 vs 0.07) on CRAFTtest; and ROUGE-1 (0.33 vs 0.28), ROUGE-2 (0.11 vs 0.07), and ROUGE-SU4 (0.14 vs 0.1). The word cloud visualization provides a concept-oriented summary of the text and allows users to retrieve salient content

according to their specific interests and requirements. The system can be used for summarizing and identifying relevant content from full-text articles from a variety of information sources such as Medline, Cochrane, UpToDate (http://www.uptodate.com/), and from clinical text such as clinical notes, radiology reports, etc. The system incorporates several features to address the challenges related with extracting information from large volumes of text. Future work will focus on extrinsic evaluation of the system in both patient care and research settings.

CHAPTER 3

SEACOIN 2.0 – AN INTERACTIVE MINING AND VISUALIZATION TOOL FOR INFORMATION RETRIEVAL, SUMMARIZATION AND KNOWLEDGE DISCOVERY

ABSTRACT

Objective: The rapidly increasing size of biomedical databases such as Medline requires use of intelligent data mining methods for information extraction and summarization. Existing biomedical text-mining tools have limited capabilities for inferring topological and network relationships between biomedical terms. Very often too much is returned during summarization leading to information overload. Furthermore, literature based discoveries could be hard to interpret if the network is too complex.

Methods: SEACOIN2.0 generates k-ary relational networks of biomedical terms using a novel association rule mining algorithm to facilitate efficient information retrieval, summarization, and visual data exploration. Multi-level k-ary trees are used for "drill-down" summarization and hypothesis generation. Summarization is presented via multiple dynamic visualization panels and an interactive word cloud. LexRank algorithm is used to identify salient sentences in top abstracts related to the query. We evaluate the system performance in information retrieval and relation extraction using the BioCreative 2013 Track 3 learning corpus.

Results: An average F-measure of 94% was achieved for document retrieval and an average precision of 88% was achieved for identification of top co-occurrence terms. The system allows interactive mining of complex implicit and explicit relationships among biomedical entities (genes, chemicals, diseases/disorders, mutations, etc.) and provides a

framework for including literature based PheWAS as demonstrated by replication of results from a recently published EMR based PheWAS.

Conclusion: We present herein SEACOIN2.0, an interactive visual mining tool for graph summarization of Medline abstracts and literature based discovery. SEACOIN2.0 addresses the problem of "information overload" and can help clinicians and biomedical researchers meet their information needs. The system facilitates literature based PheWAS (Lit-PheWAS).

Availability and Implementation: The system is available at: https://newmedicalor.isye.gatech.edu/SEACOIN2/

3.1 Introduction

Complex diseases such as cancer, diabetes, and cardiomyopathy involve multilevel interactions of cellular processes (Dyugu 2014). Knowledge and discovery of the multitude of interactions and relationships between genes, proteins, metabolites, mutations, epigenetic modifications and environmental factors is essential for understanding the pathophysiology of diseases (Aebersold 2008, Dyugu 2014). Rapid developments in biomedical research over the last two decades have enhanced our ability to study these relationships, which has led to an exponential growth in information available in the form of scientific literature, experimental datasets, and publicly available biomedical databases (Shatkay 2005, Faro 2011).

Currently, over 25 million articles are available in PubMed (http://www.ncbi.nlm.nih.gov/pubmed). In addition to scientific literature, structured and unstructured clinical text information is also recorded in electronic health record systems, which are now being used for phenotyping patients for genetic studies and for epidemic

detection (Denny 2013). Studies have shown that it could take up to five years to keep up with the articles that are published per day (Scherf 2005). To further scientific advances, intelligent tools and algorithms are required for rapid processing of the vast amount of free text available in numerous heterogeneous sources.

Biomedical text mining is a growing research area that involves information retrieval, information extraction, named entity recognition, knowledge discovery, and summarization from scientific literature (Rebholz-Schuhmann 2012, Cohen 2013). The concept of literature based discovery was established by Swanson based on his findings of the implicit relationship between fish-oil and Raynaud's disease (Swanson 1986). Various text-mining tools have been developed since then for inferring gene-gene relationships (Liu 2006), identifying semantic relations (Sahay 2006, Zhang 2011), gene-drug interactions (Griffith 2013), drug-disease relationships (Qu 2009), gene-chemical-disease relationships (Davis 2009), gene-disease interactions (Kim 2013), protein-protein interactions (Fernandez 2007, Kwon 2014), drug repurposing (Andronis 2010), and automated hypothesis generation (Liekens 2013). However, the ability to summarize the large body of scientific literature and obtain a systems level overview of the interactions without running into the problem of "information overload" still remains a challenge.

In a recent review, Lu et al. surveyed 28 text-mining tools for searching biomedical literature (Lu 2011). The authors categorized the existing systems based on features such as ranking search results, clustering search results into topics, extracting and displaying semantics and relations, and improving search interface and retrieval experience. Only 2 out of the 28 tools reviewed had graph-based visualization

capabilities. Visual data mining facilitates exploration of large volumes of data by combining data mining methods with information visualization techniques. Previous studies have shown that the process of data exploration can be made more effective by including the human knowledge in the exploration process (Keim 2002). Although application of graph-based methods for summarization has been previously demonstrated, the complexity and size of the graphs increases as the size of the documents and particularly becomes challenging and hard to interpret for more than 500 citations (Zhang 2009, Preiss 2015).

Our group has previously developed the SEACOIN system (SEACOIN, Lee 2011), **S**earch **E**xplore Analyze **CO**nnect **IN**spire, as a proof-of-principle tool to address the problem of "information overload" by utilizing multi-level k-ary trees for summarization. SEACOIN was designed to various needs of biomedical researchers related to literature mining. The system includes features such as depth of retrieved information, familiar and simple-to-use interface, information overload, number of returned pages, and design of result page format that is conducive to users' understanding of results. Its major summarization features include efficient information retrieval via navigational relationship networks and one-page summarization of abstracts related to a query. The system performance was compared with two other systems, Anne-O-Tate (Smalheiser 2008) and Also Try (Lu 2009); and it was shown that SEACOIN provided most gradual and consistent filtering (Lee 2011).

In this paper, we introduce SEACOIN2.0, an improved system with a re-designed text mining framework and various new features as compared to the previous version: a) usage of a association mining algorithm that incorporates the "salience" criteria based on

pointwise mutual information (Wren 2004, Esteban 2009), TF-IDF, and number of matching documents for ranking associated terms; b) incorporation of a controlled vocabulary generated using PubTator (Wei 2013), MeSH (Coletti 2001), and SNOMED-CT to restrict the terms in the k-ary relation trees to genes, proteins, biological processes, chemicals, species, mutations, microRNAs, histone modifications, diseases and disorders; c) incorporation of "open" discovery association mining for "drill-down" literature based discovery including literature-based PheWAS; d) an improved text preprocessing workflow that incorporates stemming and stop words removal; and e) extractive summarization of the top 30 abstracts using the MINTS algorithm (Chapter 2).

The key advances of the current research over existing text-mining tools include the usage of k-ary tree structures based on a salience criteria for query-expansion to enhance information retrieval and visualization of multi-level interactions across biomedical entities in an organized and hierarchical manner. The sub-trees in the network allow users to discover and visualize association of the query term(s) with different biomedical entities, biological processes, environmental exposures, and diseases. The software provides an interactive interface to enhance information retrieval and facilitates discovery of implicit and explicit associations between biomedical entities for hypothesis generation. We demonstrate the utility of the system for information extraction and knowledge discovery using an annotated corpus of 1,112 PubMed abstracts that was used in the BioCreative IV Track 3 task (Arighi 2014). The corpus includes annotations for diseases, genes, proteins, chemicals, and action terms associated with each abstract. We introduce the concept of "literature based PheWAS" in this work, and replicate the results

from a recently published PheWAS to demonstrate the utility of the drill-down association-based approach presented here for discovering novel associations.

3.2 SYSTEM AND METHODS

SEACOIN2.0 is developed using PHP, Adobe FLEX, Java, MALLET (McCallum 2002), Apache Lucene, RaVis Flex library (https://code.google.com/p/birdeye/wiki/RaVis), and MySQL (Figure 3.1). The description of the methods and implementation of different components of the system is provided below.

3.2.1 Medline data

An indexed database of about 14 million abstracts from all articles in the Medline database published between 1975 and 2014 was created using Perl, Entrez e-utilities tools, and Apache Lucene.

3.2.2 Controlled vocabulary dictionary and stop words

First, a term-frequency distribution was generated for ~3.2 million words found in the 14 million abstracts. Each word in this list was then evaluated using databases like MeSH (Coletti 2001), Snomed CT (disorders, findings) and the PubTator database (Wei 2013; a database of includes annotated list of genes, diseases, chemicals, mutations, and species in PubMed), and the following databases in NCBI: Entrez Gene, Protein, PubChem compounds and substances, SNP, Epigenomics, and Taxonomy (NCBI Resource Coordinators 2013).

Additionally, the WordNet database (Miller 1995, Fellbaum 1998) was used to filter words that did not overlap with any terms in PubTator and were not classified under the following categories: noun.state, noun.phenomenon, noun.artifact, noun.process, noun.animal, noun.body, noun.substance, and verb.body.

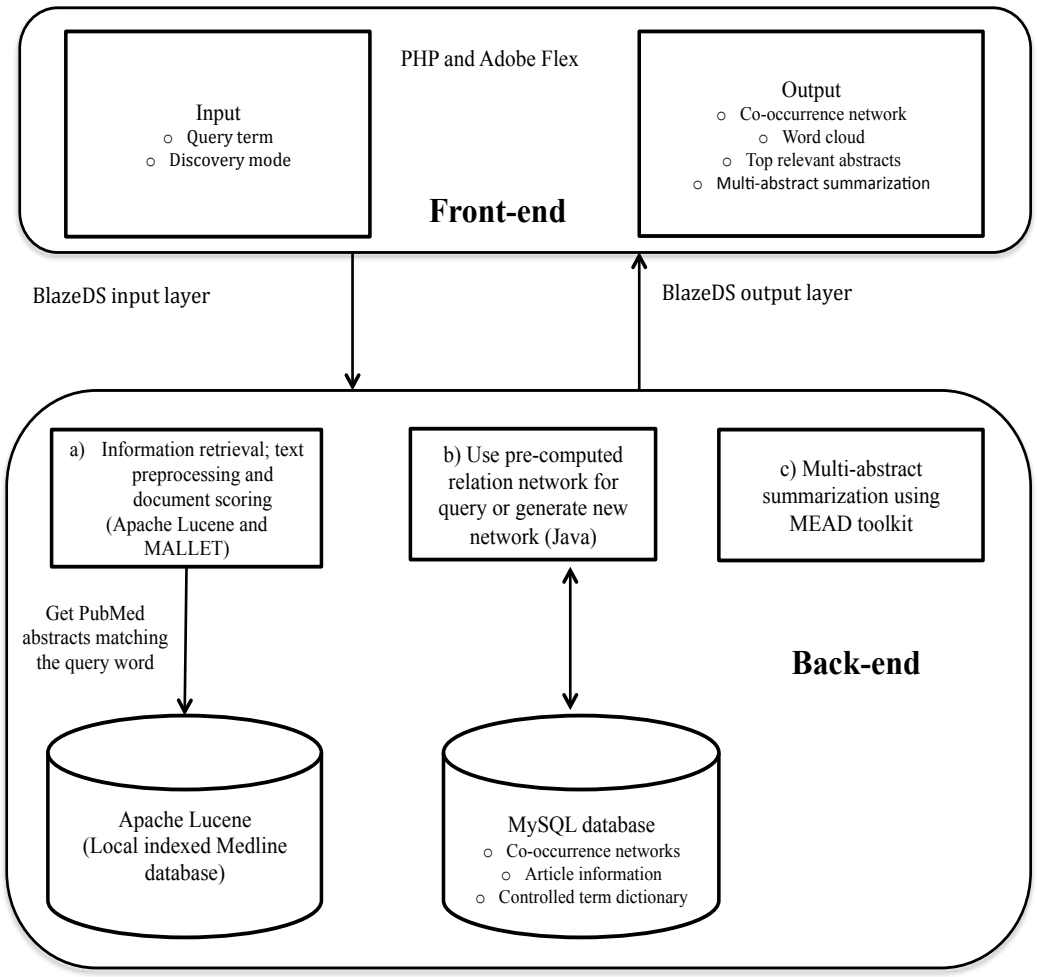

Figure 3.1. SEACOIN2.0 System Architecture

3.2.3 Text preprocessing

All abstracts were preprocessed in two stages: a) tokenization; b) filtering based on controlled vocabulary and stop words criteria (Figure 3.2).

a) Tokenization

Word tokenization involves segmentation of text into words, punctuations, whitespaces, etc. (Jensen 2006). A dictionary of over 3 million unique words found in the Medline abstracts was generated.

b) Filtering based on controlled vocabulary and stop words:

A word probability distribution using the entire collection of PubMed abstracts published between 1975 and 2015 was generated. The top most commonly occurring words ($p>0.005$) along with the stop words used by PubMed (http://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T43) were removed from the abstracts. Additionally, words that were not included in the controlled word dictionary (as described in 2.2) were also filtered from the abstracts.

3.2.4 Document indexing, searching, and ranking

In order to facilitate efficient document retrieval during various text-mining stages, the preprocessed PubMed abstracts are stored into an indexed database using the Apache Lucene package (URL: http://lucene.apache.org/). Lucene is an open-source Java based text search-engine library that includes highly efficient search and document scoring algorithms that allow querying based on phrases, wild cards and Boolean operators. Lucene scores the documents based on the term frequency-inverse document frequency (TF-IDF) scoring scheme as described in the software documentation available at: http://lucene.apache.org/core/3_6_0/api/core/org/apache/lucene/search/Similarity.html. The indexed documents are scored according to their relevance with respect to the query term, t_1.

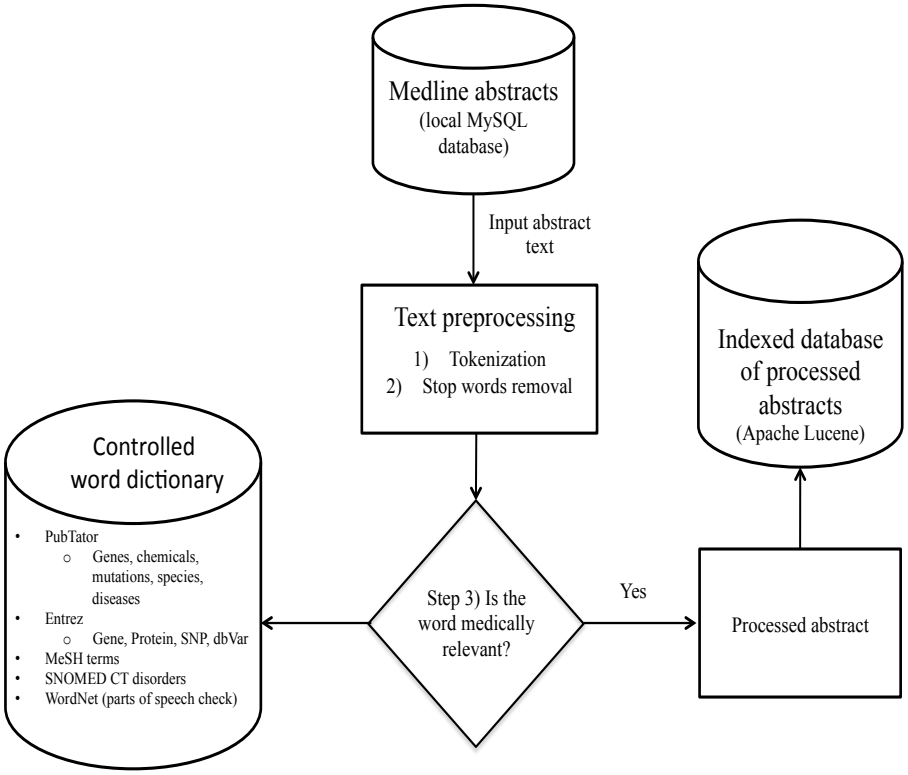

Figure 3.2. Pre-processing workflow for abstracts

3.2.5 Statistical association analysis and k-ary trees

A k-ary tree is a rooted tree where each parent node has up to k child nodes. The maximum number of nodes, N, in a k-ary tree with h levels is given by

$$N = (k^{(h+1)} - 1)/(k - 1) \qquad (1)$$

The system generates a 5-ary tree for each query term with h={1,2,3} levels. At each level, the information retrieval process within the system (Section 2.4) is performed in either open or closed discovery mode (Figure 3; Andronis 2010). In the closed discovery mode, abstracts containing all ancestors (e.g.: A and B at level 2 in Figure 3a) in the query term (AB) are used to find the child nodes. In the open discovery mode, the abstracts containing the most immediate predecessor term (e.g.: only term B at level 2 in Figure 3b) in the query are returned. This allows discovery of both implicit and explicit

relationships between biomedical entities and generate novel hypothesis as shown in Figure 3 for a 1-ary tree scenario. In the open discovery mode (Figure 3b), only the most immediate parent node is used for document searching and association mining as Top five co-occurring words are determined for each query based on the salience criteria and a k-ary co-occurrence network with 3 levels (e.g.: hypertension -> angiotensin -> pericytes -> Nitroarginine) is generated.

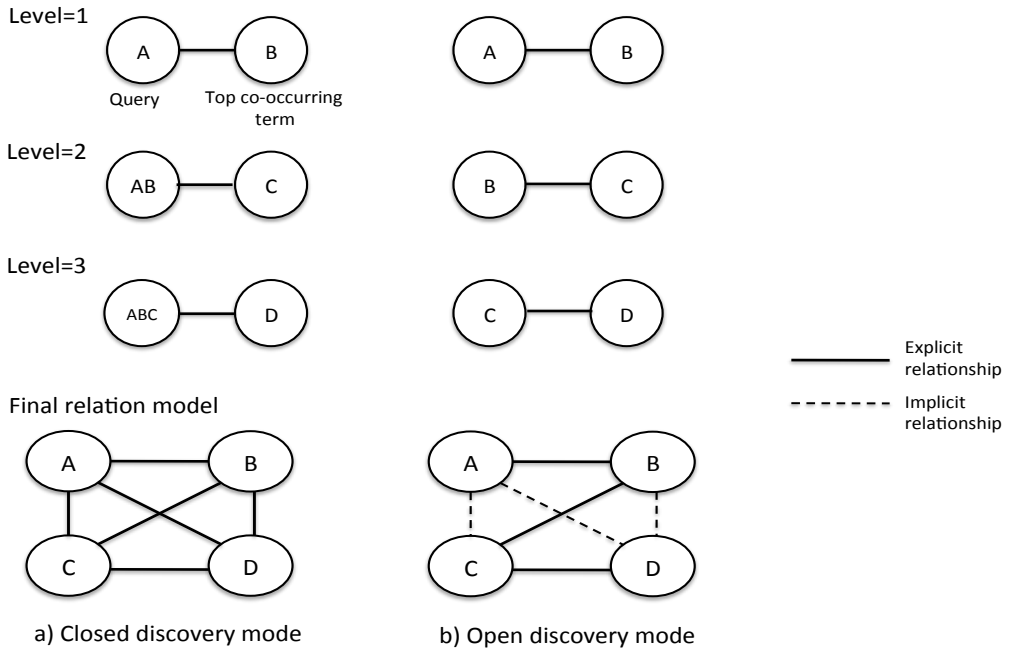

Figure 3.3. Illustration of the closed (a) and open (b) discovery modes for a 1-ary relation tree with three levels

A combined saliency criteria based on pointwise mutual information (PMI; Wren 2004, Esteban 2009) and TF-IDF is used to generate an ordered tree using the top co-occurring terms at each node (query term) as described below:

PMI between the query term (t_1) and term t_2 is computed using the formula in (2). If the probability of association of the two terms is greater than by chance, then a tf-idf criterion is used to assign the association score (3).

$$PMI(t_1, t_2) = i * \log_2 \left(\frac{p(t_1 \text{ and } t_2)}{p(t_1)\, p(t_2)} \right) \qquad (2)$$

where

i is a Boolean variable {0,1} such that i=1 if term t_2 is present in the controlled vocabulary, 0 otherwise;

$p(t_1)$ is the probability of term 1 in the corpus, $p(t_2)$ is the probability of term 2 in the corpus, and $p(t_1 \text{ and } t_2)$ is the probability of co-occurrence of terms 1 and 2 in the corpus.

If PMI>1.5, then

$$score(t_1, t_2) = \log_2(tf(t_2,t_1)) * IDF(t_2) \qquad (3)$$

where

$tf(t_2, t_1)$= frequency of term t_2 in documents related to term t_1, $D_{t1,}$ and

$IDF(t_2)=1 + \log_2 \left(\frac{\#\text{ of documents}}{\#\text{ of documents with term}} \right)$

else

$score(t_1, t_2) = 0$

In the closed discovery mode, the same procedure is repeated at levels {h=2,3} in the tree, but using equations (4) and (5)

$$PMI(t_1,t_2,\ldots,t_m) = \log_2 \left(\frac{p(t_1 \text{ and } t_2 \text{ and } \ldots t_m)}{p(t_1)\, p(t_2) \ldots p(t_m)} \right) \qquad (4)$$

$$score(t_1, t_2,\ldots,t_m) = \log_2(tf(t_m,\ldots,t_2,t_1)) * IDF(t_m) \qquad (5)$$

This multi-step approach reduces the risk of incorporating random co-occurrences with commonly used words in the relation networks (Esteban 2009).

3.2.6 Drill-down summarization and hypothesis generation

The k-ary trees allow systematic exploration and visual mining of the association pattern of co-occurring terms. The tree structure extends the original ABC model to multiple levels, thereby making the hypothesis generation process more efficient while preventing information overload. The user interface includes a scroll bar that allows users to control the number of levels in the k-ary tree, h={1,2,3} as illustrated in Figure 3.4. This feature provides improved navigation of the relation trees as opposed to static graphs and facilitates efficient visual mining of multi-level relations and hypothesis generation. Futhermore, exploring the network at deeper levels enhances the ability to discover "hidden" knowledge, which is not feasible using single level association analysis. Figure 3.4 shows the network relationship of glutathione with genes, chemicals, enzymes, exposures, biological processes, and diseases/disorders generated using SEACOIN2.0 in the closed discovery mode. Figure 3.4a shows the 5-ary relational network for glutathione with one level. The network provides an overview of the associations of glutathione with enzymes (reductase, transferase, peroxidase), and cysteine and disulfide. Figure 3.4b shows level 2 associations using "glutathione AND transferase", "glutathione AND reductase", "glutathione AND peroxidase", etc. Figure 3.4c shows level 3 associations using "glutathione" and terms from levels 1 and 2 to formulate the k-ary search query. Figure 3.4d shows a sub-network for query "glutathione AND transferease AND cancer".

3.2.7 Interactive Word/Topic Cloud:

An interactive word cloud (Lee 2011) is generated for the top 30 terms selected based on the association rule learning algorithm described above. The size of the terms in the word cloud indicates the ranking. Users can click on the terms to update the list of retrieved Medline abstracts that contain both the query term and selected term. This facilitates automated query expansion and document filtering.

3.2.8 Query-based extractive summarization

Summarization is the task of representing the information in the original text in fewer words (Cohen 2013). LexRank is a graph-based summarization method that uses cosine similarity and eigenvector centrality to determine the relevance of individual sentences (Erkan 2004). The highest scoring sentence is assigned a score of 1. Our system uses the LexRank algorithm implemented in the PERL MEAD toolkit (http://www.summarization.com/mead/) for identifying most relevant sentences from the top 30 abstracts related to the query. Only the sentences that include at least two words from the controlled vocabulary are included in the summarization process.

3.2.9 MySQL database

The SEACOIN MySQL database includes the PubMed entries (PubMed IDs, Title, Abstracts, Authors, Affiliations, Citation counts in PubMed Central), MeSH entries (MeSH terms and concepts), PubTator terms, stop words, WordNet 2.0 database, and previous queried computed co-occurrence networks (described in Section 3.2.6).

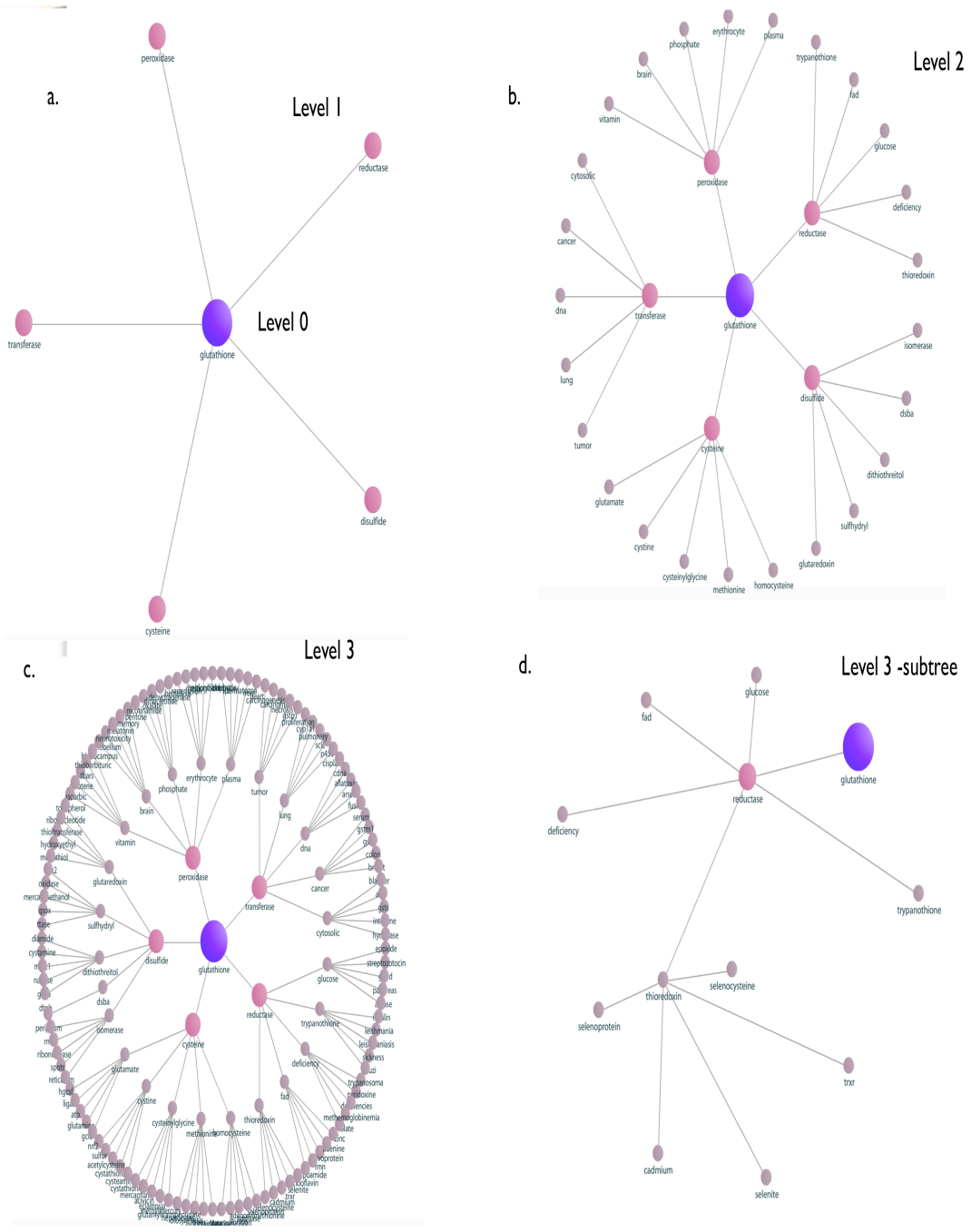

Figure 3.4. Multi-level summarization of abstracts related to "glutathione" in closed discovery mode. a) level 1: top 5 salient terms with glutathione; b) level 2: top salient terms in abstracts related to "glutathione AND reductase", "glutathione AND cysteine", "glutathione AND transferase", "glutathione AND disulfide", and "glutathione AND peroxidase"; c) level 3: top salient terms in abstracts related to "glutathione" and salient terms from levels 1 and 2; d) users can dynamically adjust the size and shape of the tree to reduce complexity.

3.2.10 Web interface

All features are embedded within a web-based system (https://newmedicalor.isye.gatech.edu/SEACOIN2/). Figure 3.5 shows schematically the information flow process within SEACOIN. Specifically, the system takes as input a query term that triggers the search process to:

- Retrieve relevant abstracts according to the open or closed discovery mode criteria from the preprocessed indexed database,
- Process the relevant abstracts to generate a k-ary network,
- Generate a word cloud based on the top 20 co-occurring terms based on the query,
- Summarize the content of up to top 30 most relevant abstracts related to the query

These features combined with the features in the original version of SEACOIN can be used to address various scientific queries as illustrated in Section 3.

3.3. Results and Evaluation

The annotated set of 1,112 abstracts from the BioCreative IV Track 3 CTD learning corpus was used to evaluate the performance of the system using standard text mining metrics such as precision, recall, and balanced F-measure (Cohen 2013). The corpus includes curated annotations for each abstract such as chemical names, gene names, disease names, and action terms (Arighi 2014). Each abstract in the corpus was preprocessed using the workflow outlined in Section 2.2 and indexed using Apache Lucene. Co-occurrence networks were generated as described in Sections 2.3 and 2.4. The following five biomedical terms were used for evaluating the document retrieval and

relationship extraction: hypertension, schizophrenia, myocardial infarction, trpv1, and cocaine (Table 3.1 and 3.2).

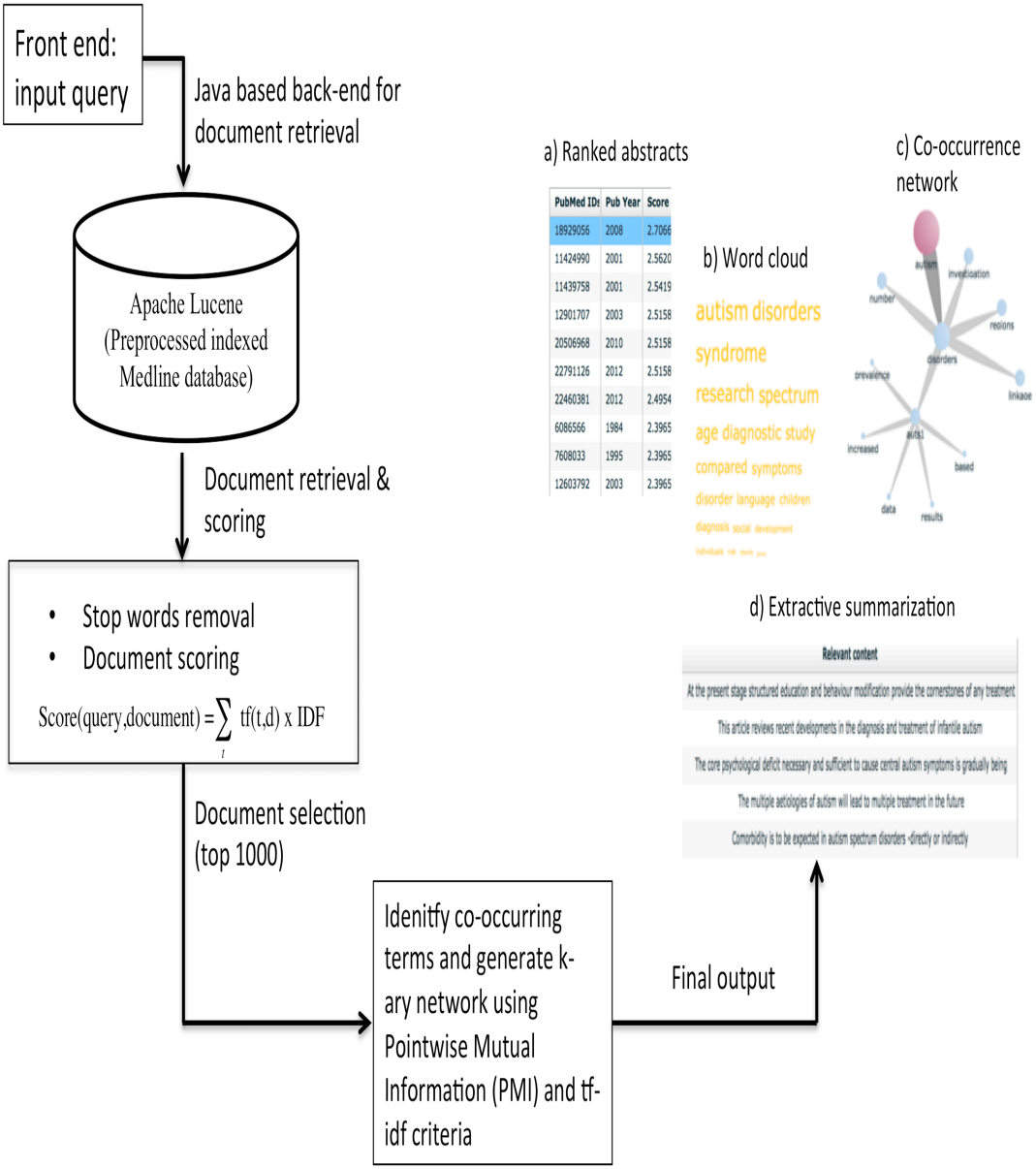

Figure 3.5. Information flow diagram

3.3.1 Document retrieval evaluation

Each abstract in the gold-standard corpus was annotated with genes, diseases, and chemicals. On an average, a precision of 89%, a recall of 100%, and a F-measure of 94% is achieved for document retrieval (Table 3.1).

Table 3.1. SEACOIN 2.0 document retrieval evaluation in terms of precision, recall and F-measure

|S|	Query	Precision	Recall	F-measure
1	Hypertension	0.897	1	0.947
2	Schizophrenia	0.727	1	0.842
3	Myocardial Infarction	0.904	1	0.95
4	trpv1	1	1	1
5.	Cocaine	1	1	1
	Average	0.89	1	0.94

3.3.2 Relationship extraction evaluation

An average precision of 88% is achieved for the top 10 co-occurring terms with the query term (Table 2), respectively. The list of terms found by SEACOIN but not present in the curated annotation list include names of enzymes (hce1), physical state (hyperactivity) and anatomical structures (amygdala), demonstrating the ability of the system to detect biologically relevant information in an unsupervised fashion without manual intervention.

Table 3.2. Evaluation of query-term relationship extraction using SEACOIN 2.0 in terms of precision. The [*] symbol represents terms that did not overlap with the curated annotations.

Query	Precision (%)	List of words extracted by SEACOIN2.0
Hypertension	90	Aliskiren, Atenolol, Losartan, Ouabain, Felodipine, Acarbose, Ramipril, hspd1, Clonidine, smmlck*
Schizophrenia	100	acp103, norZTP, bmy14802, SSR181507, Risperidone, Clozapine, Olanzapine, fkbp51, cnr1, bl1020
Myocardial Infarction	90	Epinephrine*, Disopyramide, Sevoflurane, Serca, Tetrandrine, Betaine, Telmisartan, Caffeine, Cinaciguat (BAY 58-2667), Isoprenaline/Isoproterenol
trpv1	90	AMG9810, Capsaicin, Drimanal, Hyperalgesia, Polygodial, Morphine, Capsazepine, drrs*, cgrp/calc1, sb366791
Cocaine	70	Hce1*, Amphetamine, Desipramine, Amygdala*, Minocycline, ga2-50, darpp32, Lidocaine, mpfc*, mecp2
Average	88	

3.3.3 Hypothesis generation evaluation using BioCorpus

The search for "schizophrenia" in the evaluation corpus of 1,112 in the open discovery mode identified 11 abstracts related to schizophrenia and one of the subtrees includes schizophrenia (level 0) -> bmy14802 (level 1) -> dopa (level 2) -> fosb (level 3) as shown in Figure 3.6a. As shown in Figure 3.6b, "fosb" was not mentioned in any of the abstracts matching "schizophrenia" or "BMY14802", an anti-psychotic drug. Therefore, it can be hypothesized that the fosb gene and schizophrenia are possibly linked since our knowledgebase is restricted to the abstracts in the evaluation corpus.

An independent PubMed search using "schizophrenia" and "fosb" resulted in 10 hits (at the time of this writing). One of the articles supported the hypothesis that schizophrenia, fosb and BMY14802 are linked. In their recently published work, Dietz et al. showed that the fosb expression levels increased in schizophrenic patients who were prescribed antipsychotic drugs, while no effect in fosb levels was observed in patients who were medication free (Dietz 2014).

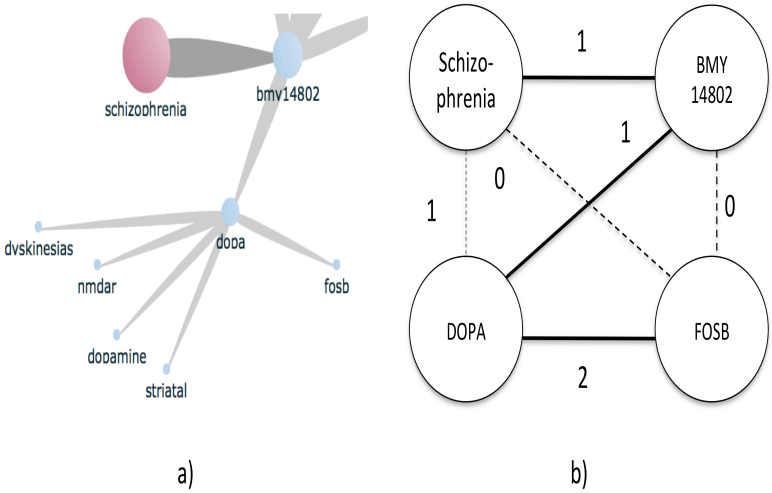

Figure 3.6. Illustration of literature-based discovery based on the evaluation corpus for "schizophrenia" a) A subtree/branch from the relation tree for "schizophrenia" generated bySEACOIN2.0 shows the relationship between schizophrenia, BMY14802, dopa, and fosb; b) Conceptual representation of the implicit (dotted lines) and explicit (bold lines) relationships in

a). The numbers on the edges correspond to the number of abstracts containing the connected terms (nodes).

3.3.4 Replication of Swanson's fish oil <-> viscosity <-> Raynaud disease discovery

In 1986, Swanson discovered an implicit relationship between fish oil and Raynaud disease based on the common link between the two, viscosity (Swanson 1986). We restricted the Medline search to only articles published before 1986 and conducted a search for "fish oil" in open discovery mode. One of the sub-trees in the network linked "fish oil" <-> "maxepa" <-> "viscosity" <-> "Raynaud" (Figure 3.7). A connection hypothesized by Swanson, which was later experimentally validated.

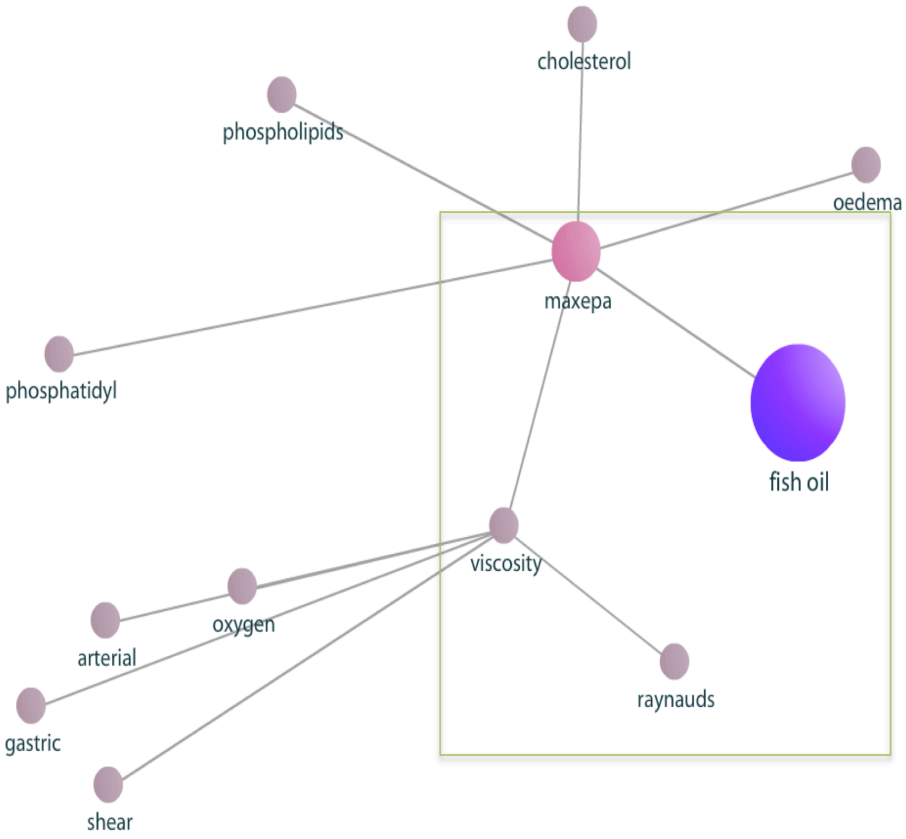

Figure 3.7. Open discovery summarization graph for "fish oil" as parent query using Medline abstracts from articles published before 1986. SEACOIN2.0 replicated Swanson's fish oil <-> viscosity <-> Raynaud disease discovery.

3.3.5 Replication of PheWAS results

Denny et al. recently published an electronic medical records (EMR) based PheWAS analysis where they studied associations between 1,358 phenotypes and 3,144 single nucleotide polymorphisms (SNPs) in a population of 14,000 individuals (Denny 2013). The authors proposed several novel associations and reproduced previously known associations, e.g. "rs12203592" (Acitinic keratosis, nonmelanoma skin cancer, brain damage, etc.) and "rs2853676" (seborrheic keratosis, glioma), (Table 2 and Figure 3 in Denny 2013). We used SEACOIN2.0 to generate open discovery summarization graphs for SNPs. The results from SEACOIN2.0 are shown in Figure 3.8.

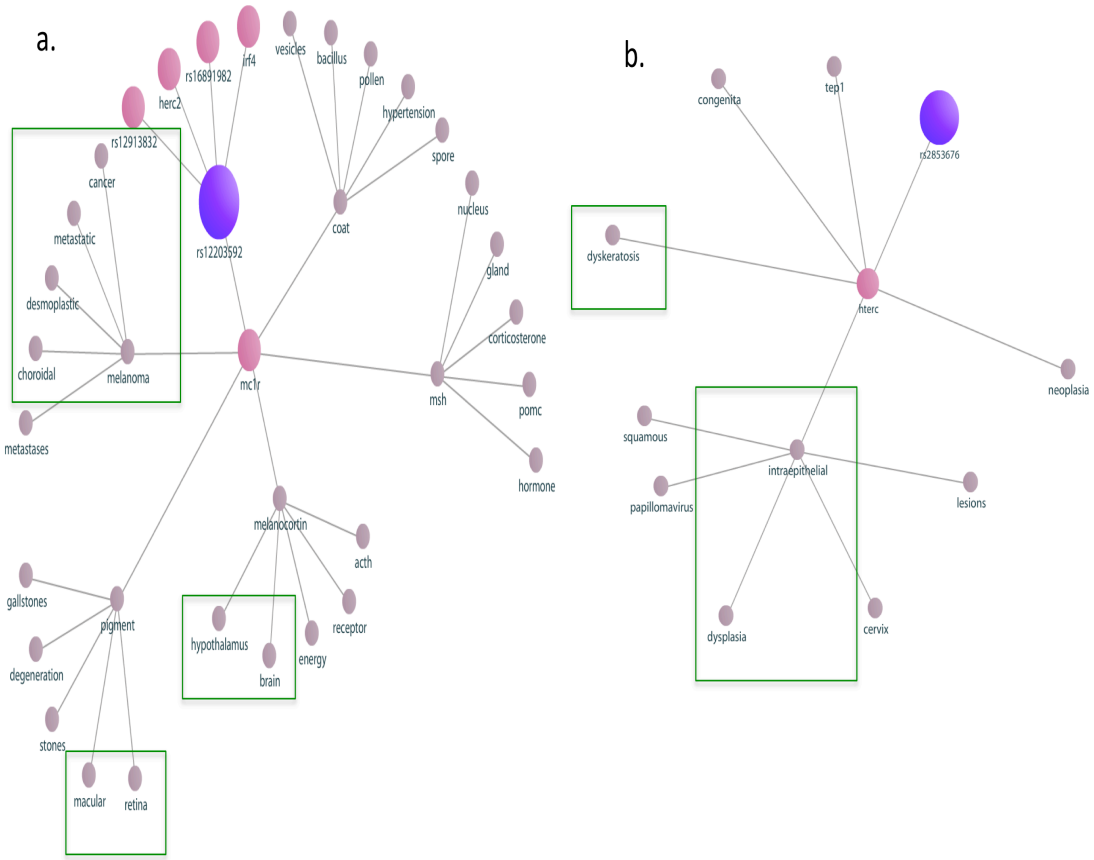

Figure 3.8. Open discovery summarization graphs for "rs12203592" and "rs2853676" using SEACOIN2.0. a) The graph suggests an association of rs12203592 with skin cancer, eye, and brain. Related associations were seen by Denny et al. although the association with brain damage was observed at p<10-3 as opposed to others which were much stronger; b) The graph suggests

an association of rs2853676 with dyskeratosis and intraepithelial dysplasia which overlap with the associations seen by Denny et al., oral mucosa and seborrheic keratosis.

The results in 3.3.3, 3.3.4, and 3.3.5 demonstrate the ability of the "drill-down" discovery algorithm proposed in this work to extract implicit relationships from literature to facilitate hypothesis generation and knowledge discovery.

3.4 DISCUSSION

Many tools have been developed that aim to extract important information from biomedical literature. We have previously developed **S**earch **E**xplore **A**nalyze **CO**nnect **IN**spire, SEACOIN, for easy and simple to understand summarization and visualization of medical literature (Lee 2011). SEACOIN is a web-based interactive tool designed using concepts and techniques from text mining, network theory, and visual data mining. It aims to provide better understanding of the biomedical literature by finding associations between biomedical terms and discovering hidden patterns in related/unrelated documents. The early system was designed taking into consideration users' preferences and concerns related to information overload, simple interface, depth of information control, number of pages returned, and computational time.

SEACOIN 2.0 is an improved version that includes new features including hypothesis generation based on both open discovery and closed discovery models and extractive summarization of top ranked abstracts that are associated with the query. The system also features technical improvements through the use of controlled vocabulary to limit the k-ary trees to biologically relevant terms. It incorporates preprocessing steps including stemming and stop words removal. It also employs point-wise mutual information and tf-idf criteria for defining query-term co-occurrence. To save

computational effort, a “history” module is developed to store previously queried computed co-occurrence networks queried by previous users. The new system also includes extractive summarization of the top abstracts.

The system provides a one-page graphical and extractive summarization of the abstracts related to the query. A multi-level interactive topological visualization is used to present the k-ary relation network where each node represents biomedical/clinical terms associated with the query. The open and closed discovery modes for generating the interactive relation networks allow users to discover and explore implicit and explicit relationships between genes, chemicals, proteins, diseases, and mutations as shown in Figure 5. Users can easily traverse the hierarchy of the network by changing the degree of separation. The system allows increase/decrease in the depth of information as users can perform real-time filtering of the network and the corresponding documents by clicking on nodes of interest in the network.

The world cloud is a visual representation of the most frequently co-occurring words with the query term. The terms included in the word cloud are representative of the search term. Users can click on any term in the cloud to update the network and the list of returned documents according to the k-ary string structure. This allows users to dynamically expand their queries and incorporate terms and topics related to the query for enhanced information retrieval. It also provides an alternative and an efficient search strategy to retrieve information as compared to those used by search engines like Google and PubMed.

This flexible structure allows retrieval of large amount of complex information in a simple easy-to understand manner. A table including the list of articles matching the

search term is also presented. Users can choose to retrieve more information about each abstract from this table.

The evaluation of document retrieval and co-occurrence network extraction results from SEACOIN2.0 with the BioCreative IV corpus shows that the system can accurately retrieve most relevant documents related to the query without generating too many false positives. It can also discover biologically meaningful query-term associations in both open and closed discovery modes. As demonstrated in Figure 5,SEACOIN2.0 has the potential to generate novel hypothesis and discover previously unknown connections between diseases, genes, chemicals, etc.

Limitations: Currently, the system only searches for top five most co-occurring terms with the query and extends the tree up to three levels. Second, individual terms in the relation trees are not mapped to concepts, which could potentially lead to ambiguity and redundancy. Third, the system uses an offline copy of the Medline database, which might not meet the information needs of all users. Fourth, currently the system generates 5-ary trees with three levels. Although this restricts the size of the tree to a manageable level, increasing the nodes and depth of the tree can improve summarization and discovery process.

Future work will focus on increasing the scalability of the system to more complex tree structures and incorporation of PubTator BioConcepts for addressing the redundancy and ambiguity issues. We expect more sophisticated relational theory will also be explored to integrate the large amount of concepts and knowledge into meaningful summary.

3.5 CONCLUSION

SEACOIN 2.0 includes improved methods for document retrieval, information extraction and summarization, and generating co-occurrence networks to discover complex relationships among biomedical terms. We demonstrate herein thatSEACOIN2.0 can be used to retrieve PubMed articles that are most relevant to a query of interest and to generate a collective summary of the previously published studies. This will facilitate discovery of direct and indirect relationships between genes, SNPs, chemicals, and diseases by means of an interactive k-ary word co-occurrence network. We replicated Swanson's "fish oil" and "Raynaud's disease" discovery. We also demonstrate the ability of SEACOIN2.0 for literature based PheWAS by replicating associations shown between SNPs and phenotypes in a recently published EMR-based PheWAS. Hence, the updated SEACOIN system will enhance the ability to extract and summarize knowledge from different articles and help generate hypothesis for future experiments, which is otherwise hard to perform using a simple PubMed search. The ability of the system to extract implicit relationships from literature to facilitate hypothesis generation and knowledge discovery is promising. Further study will be conducted along with biologists in testing some of the potential new hypotheses identified by the system. Also, the relevance scores assigned to each PubMed article will make the literature review process more efficient as this will limit the focus to a subset of articles of interest rather than thousands of articles, which may or may not be directly relevant to the query term.

CHAPTER 4

OPTSELECT: USING AGENT-BASED MODELING AND BINARY PSO TECHNIQUES FOR ENSEMBLE FEATURE SELECTION AND STABILITY ASSESSMENT

ABSTRACT

Motivation: Recent studies have shown that the ensemble feature selection approaches are essential for generating robust classifiers. Existing methods for aggregating feature lists from different methods require use of arbitrary thresholds for selecting the top ranked features and do not account for classification accuracy while selecting the optimal set. Here we present a two-stage ensemble feature selection framework for finding the optimal set of features without compromising on classification accuracy.

Methods and Results: We present herein optSelect, a multi agent-based stochastic optimization approach for nested ensemble feature selection. Stage one involves function perturbation, where ranked list of features are generated using different methods and stage two involves data perturbation, where feature selection is performed within randomly selected subsets of the training data and the optimal set of features is selected within each set using the optSelect. The agents are assigned to different behavior states and move according to a binary PSO algorithm. A multi-objective fitness function is used to evaluate the classification accuracy of the agents. We evaluate the system performance using the random probe method and using five publicly available microarray datasets. The performance of optSelect is compared with single feature selection techniques and

existing aggregation methods. The results show that the optSelect algorithm improves the classification accuracy compared to both individual and existing rank aggregation methods. The algorithm is incorporated into an R package, optSelect.

4.1 Introduction

Biomarker discovery is a key component of translational biomedical research (Guyon 2003, Lee 2007, Christin 2010). Most omics technologies measure thousands of variables (genes, metabolites, etc.) and often fall under the category of n<<p problems that are prone to model over-fitting due to large number of variables (Guyon 2003, Reunanen 2003, Cawley 2010). The large amount of feature space requires application of variable selection techniques to identify most salient variables and generate robust classifiers. This is crucial for targeted validation experiments, designing follow-up studies, and for diagnostic purposes in clinical practice.

Numerous feature selection algorithms have been developed over the last few decades (Saeys 2007, Ma 2008, Christin 2010). The feature selection methods can be classified as: filter, wrapper, and embedded (Saeys 2007). The filter methods use statistical criteria independent of the classifier to select relevant features and the selected features, e.g. $p<0.05$, are then used to build/train the model. Methods such as t-test, ANOVA, F-test, Chi-sq test, mutual information, etc. can be classified as filter methods. The wrapper methods use a search strategy to evaluate different combinations of subsets of features and select the best model based on the evaluation using a classifier such as Support Vector Machine (SVM, Vapnik 1998). Different search algorithms such as best subset, genetic algorithms, PSO, etc. can be used for finding the optimal set of features; however these methods are prone to over-fitting (Saeys 2007, Christin 2010). The

embedded methods include methods such as recursive feature elimination based on SVM, random forests (RF), Lasso, Elasticnet where the variable selection is built-in (Guyon 2002, Breiman 2001, Tibshirani 1996, Zou 2005). For instance, Lasso is a coefficient shrinkage method and uses a L1 penalty function to assign a value greater than 0 if a feature is relevant, and 0 otherwise (Saeys 2007).

Recent articles have highlighted the importance of aggregating ranking results from multiple methods to achieve a set of stable features that are likely to be reproducible in future studies (Saeys 2008, Boulesteix 2009, Abeel 2010, He 2010). The translation of basic science findings to new interventions has been limited due to irreproducible results (Boulesteix 2009, He 2010, Halsey 2015). The two main reasons of "instable" results: a) data perturbation: inconsistency in selected feature subsets due to sampling variations; b) function perturbation: different rankings of relevance from different methods (Boulesteix 2009, He 2010). Many feature selection approaches use arbitrary rank or significance thresholds to select the number of features without thorough evaluation, which could lead to suboptimal results. Moreover, different algorithms vary differently in performance depending on the distribution of the data and within-class variability (Saeys 2008, Boulesteix 2009). Here we introduce a novel nested ensemble feature selection framework, optSelect, that performs multi-objective optimization using agent-based modeling and binary particle swarm optimization (PSO) techniques to allow aggregation of results from different methods and performs nested feature selection using random subsets of training data to evaluate feature stability (Kennedy and Eberhart 1995, Chuang 2010).

4.2 System and Methods

4.2.1 optSelect: An optimization based approach for nested ensemble feature selection and aggregation

A two-stage procedure is used for finding the optimal set of features. In stage one, a ranked list of features is generated using one or more feature selection algorithms selected by the user. The user can select t.test, f.test, recursive feature elimination, random forest, wilcox.test, lasso, elasticnet (Saeys 2007, Christin 2013). The number of features selected impacts the performance of the classifiers and the predictive accuracy (Reunanbam 2003). It is essential to find the optimal set of features. Users have the option to perform sequential backward or forward selection or choose an arbitrary cutoff to select the optimal set of features. Using different ranking criteria allows function perturbation. The different variable selection methods implemented in the CMA package in R are used at this stage (Slawiski 2010). A union of the ranked lists from different methods is used as input for stage two. Users have the option to skip stage one and use all features during the optimization procedure; however some selection criteria is recommended prior to stage two to reduce the computational time and perform ensemble selection (Saeys 2007).

In stage two, the newly developed multi-objective stochastic optimization procedure based on agent-based modeling and binary framework is used to perform data perturbation and aggregation using the results from stage one.

Agent-based models involve three key elements (Macal 2010):

- Agents and their attributes/behaviors: each agent is assigned to a behavior or rule category, e.g. follows neighbors, moves randomly, etc.

- Relationships and interactions between agents: the interactions define how the agents influence and cooperate with each other,
- Interaction with the environment: the environment provides feedback to the agent about their movements (e.g. 10-fold cross-validation accuracy)

A binary PSO algorithm is used to simulate the movement of agents according to their behavior. PSO is a stochastic optimization technique based on the movement and intelligence of swarms developed by James Kennedy and Russell Eberhart in 1995. It comprises of a number of agents/particles that constitute a swarm moving around in the search space looking for the best solution determine based on a fitness evaluation function. The movement of each particle, p_i, is determined based on a velocity vector, v_i, and a position vector, x_i. In binary PSO, the position vector, x_i, has d dimensions, where d corresponds to the number of variables (genes, chemicals, etc.). And, $x_{id}=\{0,1\}$. This allows application of the binary PSO for feature selection, where a feature is selected if $x_{id}=1$. The velocity and position vectors are updated according to equations (1) and (2).

The velocity of each particle, p_i, is updated at iteration t+1 according to the equation,

$$v_i^{t+1}=v_i^t+c_1*r_1*(pbest_i-s_i^t)+c_2*r_2*(gbest_i-x_i^t) \quad (1)$$

where,

i is the current particle

c_1 and c_2 are constant learning factors to control the social influence and global influence,

r_1 and r_2 are random numbers between [0,1] interval,

pbest is the best position of the particle has experienced based on the fitness function,

gbest is the best position experienced by any particle in the swarm based on the fitness function, and

x_i is the position in iteration t.

The velocity of the particle is restricted to be in the interval [-6,6]. The position is updated according to (2) and (3) in binary PSO based on a sigmoid transformation, S, of the velocity (Figure 1),

for d in 1 to number of variables:

$$S_d=1/(1+\exp(-v_{id}^{t+1})) \quad (2)$$

$$x_{id}=1 \text{ if } S>r_3 \quad (3)$$

0 otherwise

where

x_{id}^{t} : is the position of the i^{th} agent at time t in dimension d (gene, chemical, etc.),

v_{id}^{t+1} is the updated velocity of the i^{th} agent at iteration t+1 in dimension d,

S_{id} is the sigmoid function with values between [0,1] interval for dimension d,

r_3 is a random number between [0,1] interval

In the modified binary PSO introduced in this work, users can provide a weight vector to bias the selection process based on expert knowledge or from literature. Studies have shown that incorporating prior knowledge can improve the classification accuracy (He 2010). The random number in (3) is replaced by the weight of the feature [0 to 1] to bias the selection process. A weight of 0 would mean that the feature is always selected and included in all feature subsets for evaluation.

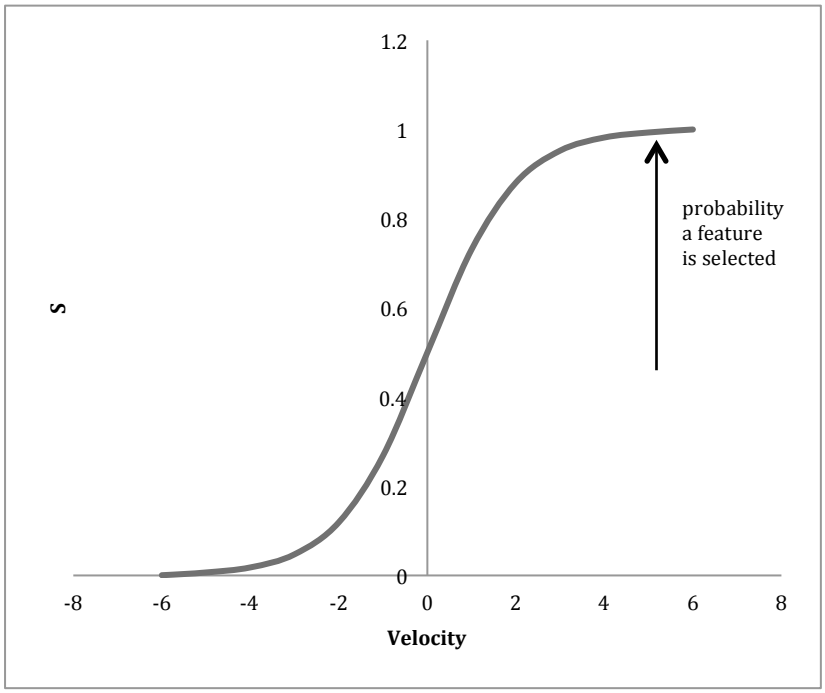

Figure 4.1. Relationship between velocity, sigmoid function, and probability of a feature being selected. The likelihood of a feature being selected increases as the velocity approaches -6 and S approaches 1.

In most existing versions of the binary PSO algorithm, all particles behave uniformly and generally follow the fully connected topology where each particle is connected to every other particle (Chuang 2008). Each agent is assigned to one of the four behavioral states {C=Confusion, S=Self-influenced, N=Influenced by nearest neighbors, G=Influenced by swarm} based on the crowd model (Wu 2014). The behavior of the current particle determines which nodes in the swarm network are chosen for interaction and updating velocities as described below:

a) Behavior=N, follows neighbors

$$v_i^{t+1}=v_i^t+c_1*r_1*(pbest_i-s_i^t)+c_2*r_2*(nbest_i-x_i^t) \quad (4)$$

where
nbest is the centroid (75^{th} percentile) of the k nearest neighbors (default k=3),

b) Behavior=G: follows global leader

$$v_i^{t+1}=v_i^t+c_1*r_1*(pbest_i-s_i^t)+c_2*r_2*(gbest_i-x_i^t) \quad (5)$$

where
gbest is the global best position in the swarm,

c) Behavior=C: moves randomly

$$v_i^{t+1}=v_i^t+c_1*r_1*(pbest_i-s_i^t)+c_2*r_2*(rpos_i-x_i^t) \quad (6)$$

where
rpos is a random position vector,

d) Behavior=S: only self-influenced

$$v_i^{t+1}=v_i^t+c_1*r_1*(pbest_i-s_i^t) \quad (7)$$

The behaviors are updated on regular intervals to prevent the population from getting stuck in local optima (user-defined parameter; default: 5 iterations). The fitness of each particle is evaluated using a nested cross-validation procedure and support vector machine (SVM) classifier as shown in Figure 2 (Vapnik 1998). The algorithm uses the internal cross-validation scheme that performs variable selection based on the training set, $train_k$, in the m-fold scheme and uses the left-out subset, $valid_m$, for evaluating the performance of the model. The internal cross-validation scheme is shown to out-perform the commonly used external cross-validation scheme where the model evaluation is performed after the selection using all samples (Cawley 2010). The search process is terminated when the distance between the centroid of the entire population and the global best agent is less than or equal to 2.

4.2.2 Evaluation experiments

Two experiments were performed to evaluate the performance of the newly developed framework. In the first experiment, a random probe evaluation was performed using the

1. Split the training set into M learning and validation sets of size Strain and Svalid (user-defined; default 80% for Strain)

2. For learning set $train_m$ and validation set $valid_m$:

Initialize the environment with N (default: 20) agents by randomly selecting features for every agent and assigning random behavior according to the behavior distribution (user-defined)

while distance(population_centroid,global_best)>1**:**

Step 1) For every particle, p, evaluate the following multi-objective function in parallel:

$$fitness_{pk} = w_1*(CVtrain_k - CVtrain_{perm}) + w_2*(CVtrain_k) + w_3*(back\_accuracy) - w_4(max\_desired\_features - current\_number\_of\_features) \quad (8)$$

where
w_1 => weight for difference between average k-fold CV vs average permuted k-fold CV from three random splits of $train_m$ (default:0.7),
w_2 => weight for 10-fold CV (default: 0.25),
w_3 => weight for back accuracy (default: 0.05),
w_4 => controlling the penalty for model complexity (default: 0.01),
back_accuracy => using a bootstrap sampling of the $valid_m$ at $fold_m$ for training and the current learning set, $train_m$, as test

Step 2) Determine the local best and global best particle that gives highest fitness

If no_change>max_no_change_iterations, then,
behavior_reset(),
position_reset()

Step 3) Update the velocity and position of every agent using equations 2-7,

Store the best of features selected for $fold_m$
Evaluate the prediction accuracy using $train_m$ for training the model and $valid_m$ as test set
Return best set of features and validation accuracy for $fold_m$

3. Calculate stability measure for each feature,

$$SM = \frac{\text{(Number of learning sets in which a feature is selected)}}{\text{Number of learning sets}}$$

4. Aggregate the results for folds 1 to M, by selecting features at different SM thresholds [0 to 1] and selecting the set with highest fitness value
5. Return the average of the accuracy of validation sets [1 to M], outer CV
6. Report results

Figure 4.2. optSelect algorithm for nested feature selection

Iris dataset (Fisher 1950). The original dataset has 4 real features and 150 instances belonging to 3 classes of the iris plant. Each real feature was scale normalized ($\mu = 0, \sigma = 1$) and 36 random features with similar distribution were included in the dataset. The instances were divided into 60% train (90; 30 per class) and 40% test (60; 20 per class). The aim was to evaluate the ability of the stage two of the optSelect algorithm to identify real features. In the second experiment, five publicly available microarray datasets: Leukemia (7129 genes, train samples=38, test samples=34, 2 classes; Golub 1999), SRBCT (2308 genes, train samples=63, test samples=25, 4 classes; Khan 2001), Prostate cancer (6033 genes, train samples=61, test samples=41, 2 classes; Singh 2002), MAQCII-ER and MAQCII-PCR (22284 genes, train samples=130, test samples=100, 2 classes) (Popovici 2010). Limma, rfe-SVM, Lasso, Elasticnet, F-test methods implemented in the CMA R package were used for feature selection in stage one. The two rank aggregation methods implemented in the rankAggreg R package, rankAggreg-Monte Carlo and rankAggreg-GA, that optimize the list of ranked features based on the Spearman's footrule were used for comparison in stage two. The evaluation process was repeated using top 5 and top 15 features from stage one. The final performance of different methods is evaluated using an independent blinded test set that was not seen by the optSelect algorithm during the model building stage. The balanced accuracy, average class-wise prediction accuracy, was used for evaluation purpose (Dreyfus 2006).

4.3 Results and Evaluation

For experiment 1, 3 out of 4 real features selected in 8 out of 10 folds. 2 were selected in all 10 folds. All random features were discarded. Both train 10-fold cross-validation accuracy and balanced accuracy for the test set were 95%.

The results for experiment 2 are summarized in tables 1 and 2. Table 1 shows the test set balanced accuracies for the test sets from the five datasets using top 5 ranked features selected by different methods (limm, rfe-SVM, Lasso, Elasticnet, F.test) in stage 1 followed by different aggregation methods in Stage 2.

Table 4.1. Evaluation using top 5 ranked features in Stage 1. The balanced accuracies for the independent test sets are reported here.

	Method\Dataset	Leukemia	SRBCT	MaqcII_ER	MaqcII_PCR	Prostate
Stage 1	Limma	0.89	0.94	0.87	0.7	0.77
	rfe-SVM	0.94	0.92	0.88	0.66	0.77
	Lasso	0.94	0.9	0.87	0.62	0.75
	Elasticnet	0.89	0.94	0.87	0.72	0.77
	F.test	0.93	0.94	0.87	0.66	0.77
Stage2 (Aggregation)	rankAggreg-monte carlo	0.89	0.92	0.87	0.74	0.77
	rankAggreg-GA	0.89	0.94	0.87	0.64	0.77
	optSelect	0.96	1	0.87	0.72	0.77

The results show that the ensemble approach improves the classification results as compared to individual methods. For both Leukemia and SRBCT test sets, the optSelect algorithm achieved highest accuracy compared to both individual methods in stage 1 and existing aggregation techniques. For the other three datasets, the optSelect algorithm gave similar performance as compared to individual and existing methods.

Table 4.2 summarizes the balanced accuracy results for the five datasets using the top 15 ranked features selected by different methods (limm, rfe-SVM, Lasso, Elasticnet, F.test) in Stage 1 followed by different aggregation methods in Stage 2.

Table 2. Evaluation using top 15 ranked features in Stage 1. The balanced accuracies for the independent test sets are reported here.

	Method\Dataset	**Leukemia**	**SRBCT**	**MaqcII_ER**	**MaqcII_PCR**	**Prostate**
Stage 1	limma	0.96	0.99	0.88	0.71	0.77
	rfe-SVM	0.96	0.661	0.86	0.68	0.77
	Lasso	0.96	1	0.87	0.64	0.77
	Elasticnet	0.96	0.66	0.88	0.68	0.77
	F.test	0.96	0.986	0.88	0.66	0.77
Stage2 (Aggregation)	rankAggreg-monte carlo	0.94	1	0.88	0.75	0.77
	rankAggreg-GA	0.96	0.96	0.87	0.73	0.77
	optSelect	0.96	1	0.9	0.73	0.83

As discussed earlier, the arbitrary thresholds for selecting top ranked features does not guarantee reproducibility of results on a test set. For instance, the performance of rfe-SVM and Elasticnet degrades by 28% for the SRBCT dataset by increasing the number of selected features to 15. On the contrary, almost all methods showed improvements in accuracy for the Leukemia dataset by using greater number of features. Overall, the

aggregation stage improved the classification accuracy with the optSelect algorithm performing comparably or better in almost all cases.

4.4 Discussion

In recent years, various articles have raised the issue of feature instability and irreproducibility that has hindered the translation of results from basic science to clinical domain (Boulesteix 2009, Abeel 2010, He 2010, Halsley 2015). Several methods for rank aggregation have been proposed (Saeys 2008). However, most of these methods use criteria independent of the classification accuracy for aggregating the results and require selection of an arbitrary threshold for top k features to determine the overlap. This could result in degradation of classification performance on an independent or unseen data set.

Here we propose a novel optimization based nested ensemble feature selection framework, optSelect, which addresses this problem by using a two-stage approach for aggregating the results. Stage one involves selection of top ranked features using different methods. The top list of features from different methods is merged and used as input for a multi agent-based optimization procedure to find the most stable set of features with good classification accuracy. The "optimal" set is determined using a multi-objective fitness function as described in Methods. The algorithm incorporates the concepts of function perturbation and data perturbation to select the most optimal and stable set of features (Boluestrix 2009, He 2010). Additionally, the algorithm is designed to prevent agents from getting stuck in local optima. Each agent interacts with other agents based on their behavior state {confusion, follows neighbors, follows global leader, self-influenced}. Search for optimal solution terminates when the global best and the centroid of the population converge, which is determined using the Euclidean distance.

The performance evaluation of the multi agent-based optimization procedure on the random probe experiment (Experiment 1) shows that the newly proposed behavior based search algorithm combined with the multi-objective classification based fitness function allows detection of relevant features even when majority of the features are randomly generate. Evaluation results for the five gene expression datasets (Experiment 2) show that the newly proposed optSelect framework allows aggregation of results from different methods without compromising for the classification accuracy. The results also highlight the dramatic changes in classification accuracies as a result of arbitrary thresholds for selecting top features. Furthermore, the performance of different independent feature selection techniques varies across different datasets. On the contrary, the optSelect algorithm performed consistently across all datasets
and improved the classification results on independent test sets in most cases. The output includes outer CV estimates, optimal set of features, and stability measures for each features. Figure 3 shows the stability measures of the 6 features from the optimal set selected by optSelect for the MAQCII-ER dataset, where the samples were classified as ER+ve vs ER-ve. The most stable feature that was reproducibly selected in all folds in the nested feature selection was estrogen receptor 1.

Limitations and future work: The current study did not assess the effect of using data normalization methods and classifiers on classification accuracy. Escalente et al. have recently showed application of PSO for full model selection (pre-processing methods, feature selection, and classification algorithms). Future work will focus on extending the current framework to full model selection.

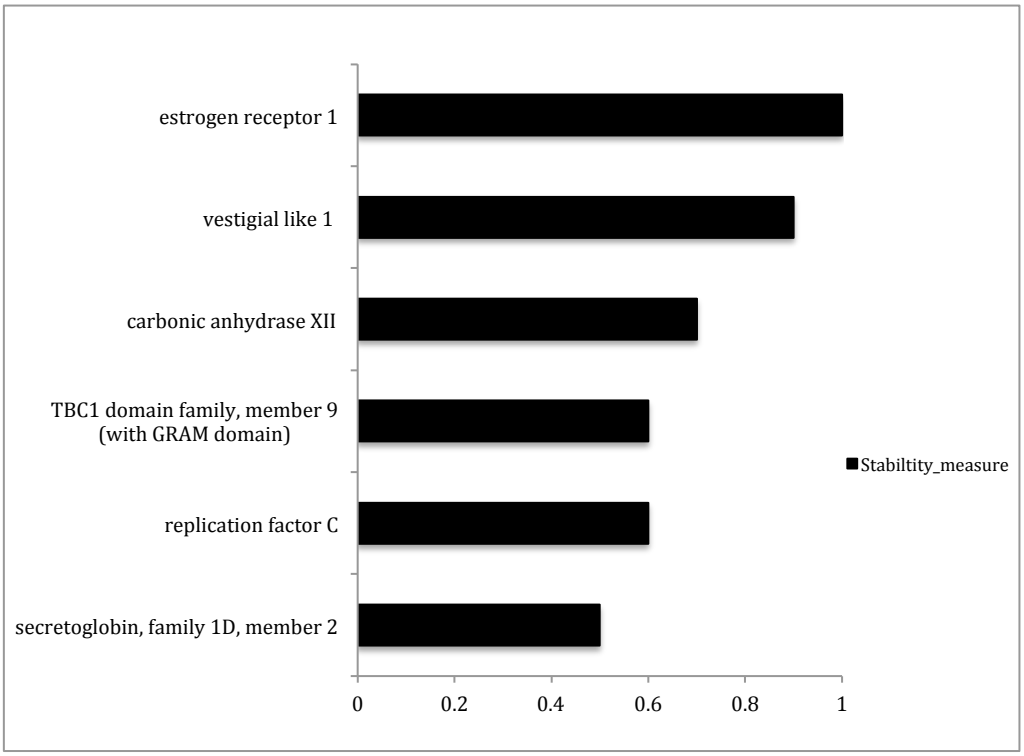

Figure 4.3. Stability measures for features in the optimal set selected by optSelect.

4.5 Conclusion

We have developed a novel multi-stage nested ensemble feature selection algorithm, optSelect, which accounts for function perturbation and data perturbation during the feature selection process. The results show that the ensemble framework improves the classification accuracy over independent feature selection methods and existing rank aggregation methods.

CHAPTER 5

APPLICATIONS

The three novel computational tools developed in this work facilitate extraction of salient information from clinical and biomedical data sources. Both text mining and feature selection tools have broad application as described in this chapter.

5.1 Clinical applications

5.1.1 Language translation systems for discharge summaries

The CoReViz system is being incorporated into an automated language translation system (Figure 5.1). The system is being developed in collaboration with Children's Hospital of Atlanta (CHOA) to generate discharge summaries in different languages for patients' with limited English proficiency (LEP).

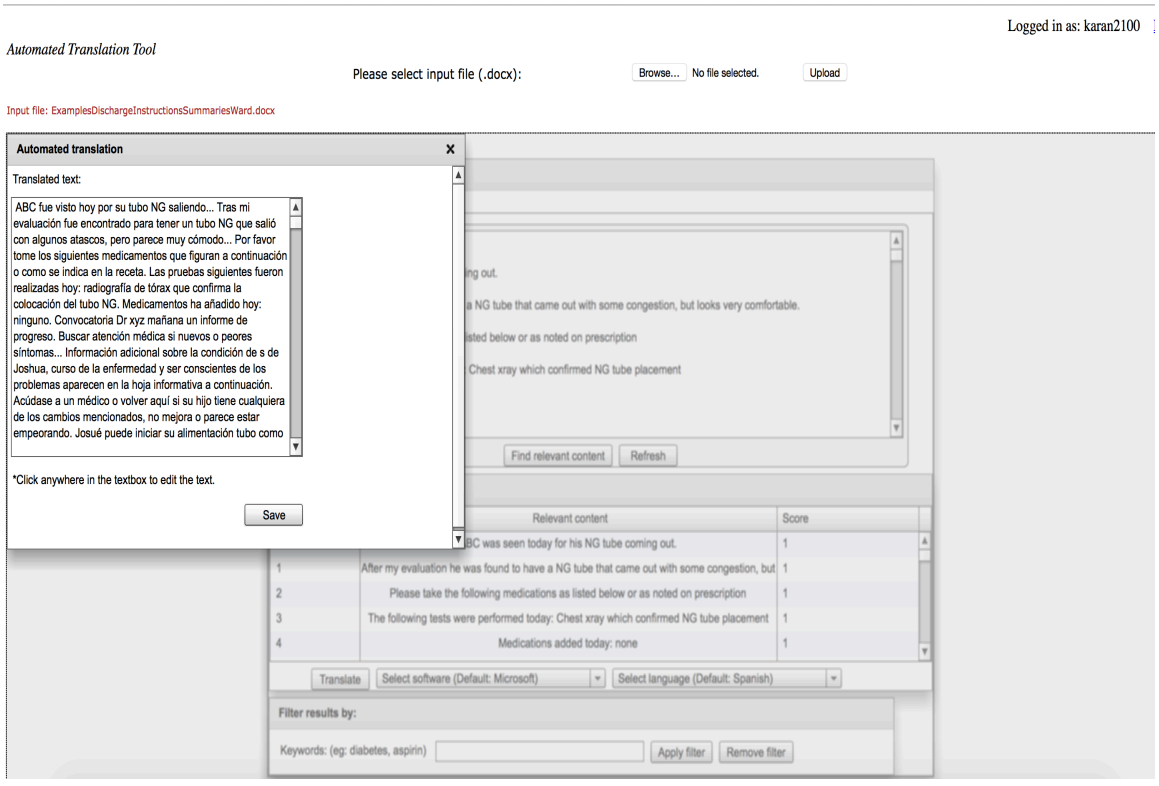

Figure 5.1 GT - Automated language translation system

(Developed by K. Uppal, Dr. Eva K. Lee in collaboration with CHOA)

The input text is first processed using the CoReViz framework presented in Chapter 2 to perform sentence segmentation and rank sentences in the input text using the MINTS algorithm. The results are then used as input for Microsoft translator to perform translation in up to 16 different international languages. The system is undergoing validation and development and will be used by clinicians at CHOA.

5.1.2 Summarization of Electronic Health Records and early detection of medical risks

As mentioned in earlier chapters, the growing amount of information in clinical databases related to patient medical history, medication records, laboratory reports, etc. is leading to the problem of information overload (Preiss 2015). In a recent publication, Feblowitz et al. proposed the AORTIS model: Aggregation, Organization, Reduction/Transformation, Interpretation, and Synthesis. All three tools developed as part of this thesis will be useful for implementing the AORTIS model in EHR systems. The framework presented in this work allows extraction and summarization of salient information from clinical text, scientific literature, omics technologies, and clinical lab measurements (Figure 5.2). For example, a search for “breast cancer” using SEACOIN2.0 identifies “estrogen” as one of the top terms in the word cloud and includes “eralpha” as one of the associated terms in the k-ary relation network (Figure 5.2). Estrogen receptor 1 gene was identified as the most stable and reproducible gene for discriminating ER positive and ER negative patients in the MAQCII breast cancer dataset using optSelect. These results indicate that the computational framework presented here can facilitate efficient and improved decision making in the clinical and biomedical domains. The three algorithms can be incorporated into the Clinical Decision Support Systems (CDSS) to help clinicians in making decisions related to diagnosis, ordering tests (if the patient history or family history indicates high risk indicators), preventing adverse drug events (if the drug being

prescribed to the patient has ingredients that can trigger allergic reactions determined based on literature mining and patient's genomics information), predictive health (if the patient's family history, medical history, exposures, and genomics information puts them under the high-risk category based on predictive modeling), etc.

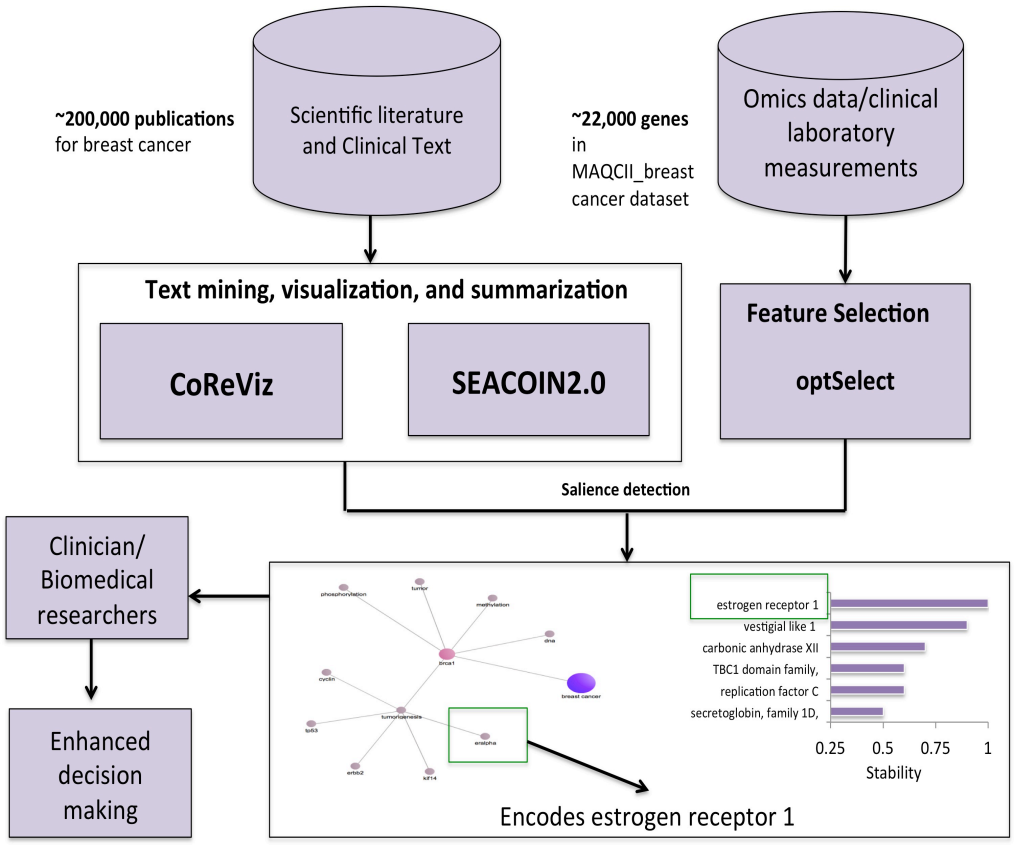

Figure 5.2. Clinical Descision Support System using CoReViz, SEACOIN2.0, and optSelect

5.2 Biomedical applications

5.2.1 Identifying discriminatory features in biomedical studies

The two-stage analysis in Chapter 4 was used to identify differentially methylated promoters using MeDIP-chip data (Koczor 2013) from 20 patient samples, ten from

Dilated Cardiomyopathy (DCM) group and the other ten from Non-Failing (NF) group. The processed data files from Nimblegen with ratio of the methylated DNA sample to the input (total DNA) sample for each DNA set relative to the peaks within promoter regions were used for analysis. A total of 19,156 unique genes were represented by at least one peak in any of the samples. However, only the genes that were present in at least 50% of the samples in either one of the classes were used to generate a m*n matrix, where m=12982 genes and n=20 samples, ten from each group. A score of 0 was assigned if a gene was not found enriched in a sample. An average relative score was used for genes represented by more than one peaks. The data was quantile normalized to address any technical and sample preparation variability across samples.

A two stage gene selection process (as was used to identify differentially methylated genes. In stage 1, differentially methylated genes were identified using two-sample Welch test, two sample Wilcoxon test, limma (Bioconductor), and SAM (samr, Bioconductor) at false discovery rate (FDR) adjusted significance level of 0.05. A gene was classified as being differentially methylated if it was selected by at least two methods. 574 genes were identified to be differentially methylated, and were then classified as hypermethyated in NF (61 genes) or hypermethylated in DCM (290 genes) if the natural log fold change (NF/DCM) was greater than 1 (or less than -1), respectively.

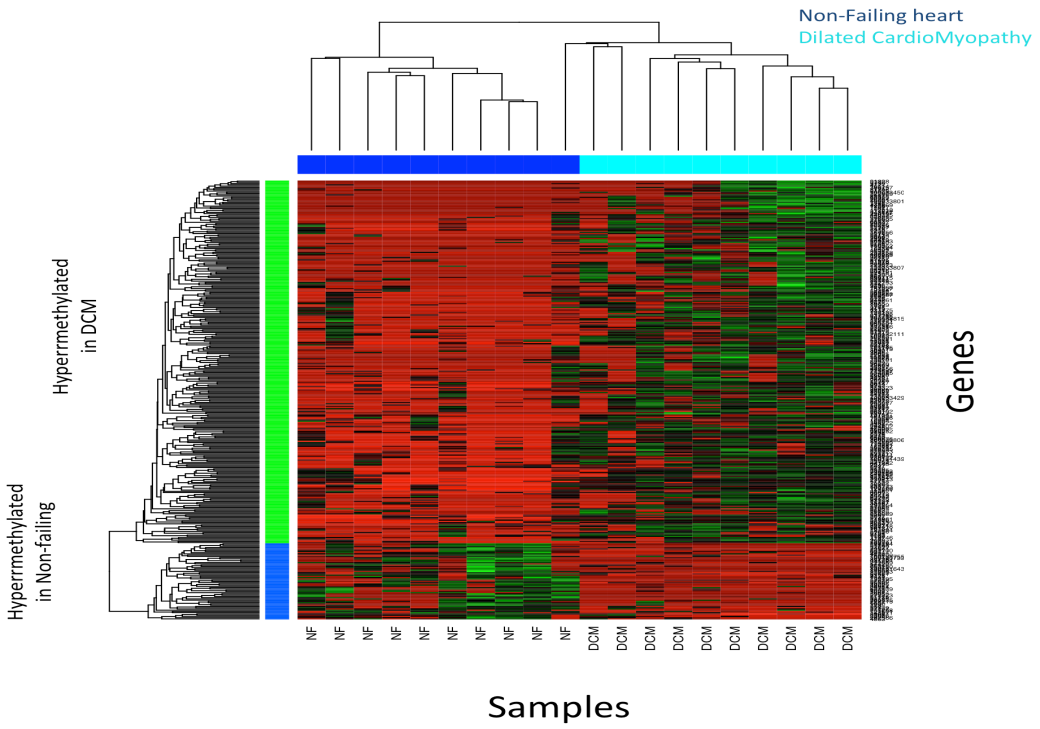

Figure 5.3 Two-way hierarchical clustering analysis using the 351 differentially methylated genes (290 hypermethylated in DCM; 61 hypermethylate in NF)

This method identified genes that are known to be involved in the cardiovascular processes and cardiomyopathy according to DAVID and DisGeNet functions annotations:

- Cardiovascular: Map2k2, RAC1, RELB, Fabp5, tbxa2r, PTGES, TNN1, myoz2, ITGB6
- Regulation of NADP pathway: NDUFAF3
- Mitochondrial genome maintenance: DNAJA3
- Cardiomyopathy: myoz2

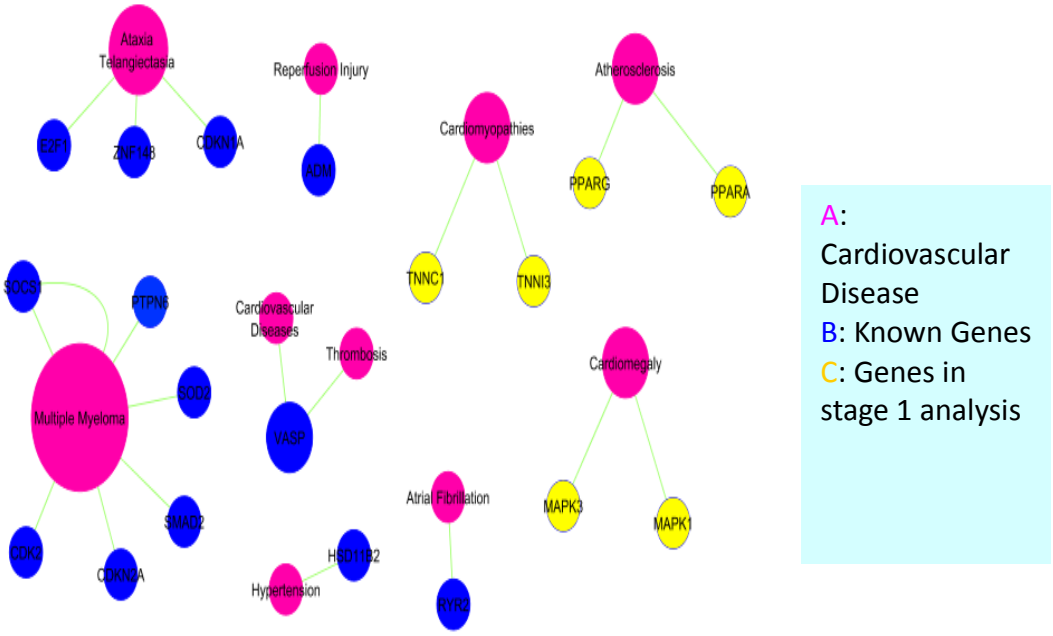

Figure 5.4 Functional validation of the differentially methylated genes identified at stage 1 that are known to be involved with cardiovascular diseases such as Cardiomyopathies, Cardiomegaly, and Atherosclerosis.

5.2.2 Identifying discriminatory sequence patterns

The optSelect framework can also be used for identifying discriminatory DNA sequence patterns between different genomic regions. The feature matrix in this scenario would correspond to the frequency of sequence patterns/l-mers, where l=length of sequence pattern as previously shown (Fletez-Brant 2013). This is particularly useful for ChIP-Seq and MeDIP-Seq techniques where the aim is to find regions enriched for specific histone modifications, transcription-factor binding sites, or CpG methylation as shown in Figure 5.5.

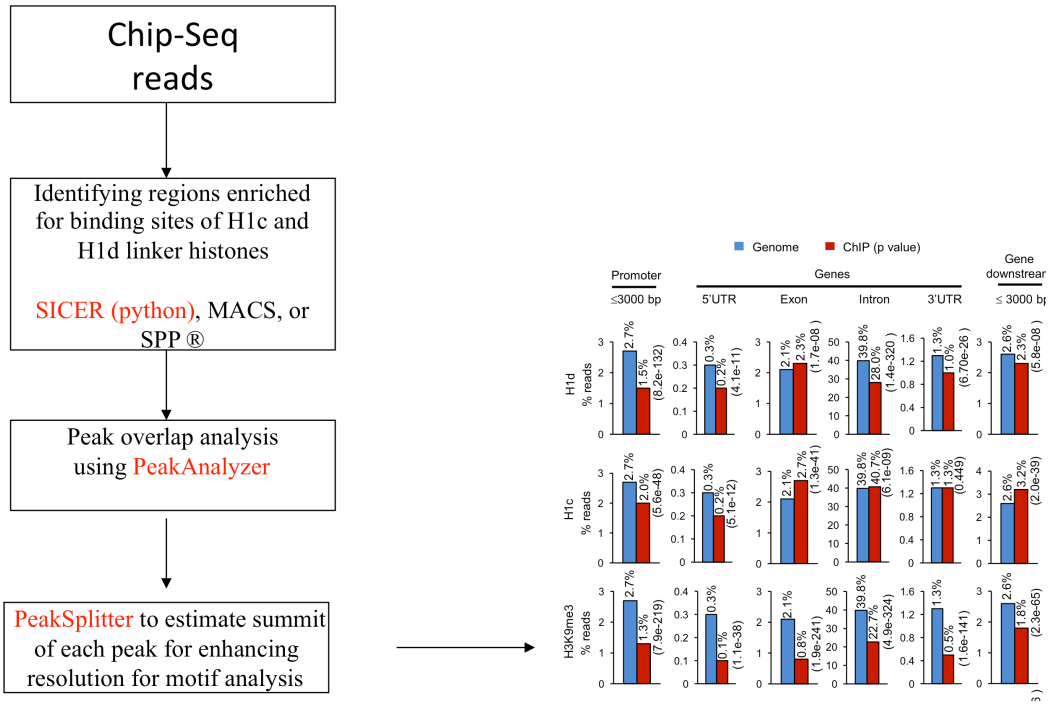

Figure 5.5. Workflow used for identifying regions enriched for linker histones and histone marks (Cao, Uppal* et al. 2013)

5.3 Closing remarks

The examples illustrated in this chapter show that the set of tools in this work can be used for different purposes in both clinical and biomedical domains.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

This thesis presents a computational framework for addressing information challenges related to precision medicine and personalized healthcare system where intelligent tools and algorithms facilitate integrating information from heterogeneous data sources. We demonstrate that algorithms based on machine learning, text mining, visualization, network theory, agent-based modeling, and optimization can facilitate detection of salient information from textual data, improve access to hidden and existing knowledge in scientific literature, and identify stable and reproducible biomarkers in case/control studies. The algorithms can be further improved and it is an active area of research. Collaborative efforts with healthcare and experimental biologists are underway and the algorithms will be improved to meet the demands of the clinical/biomedical researchers. We anticipate continued development of the three software systems/packages (CoReViz, SEACOIN2.0, and optSelect) as more and more data becomes available. The three tools can be integrated into a single intelligent system; however this will require scaling of the infrastructure to meet the computational demands of the three. Future work will focus on developing a learning system using the algorithms presented in this thesis.

REFERENCES

Aebersold et al. (2008) Report on EU-USA workshop: how systems biology can advance cancer research. Mol Oncol., 3(1), 9-17.

Andronis C, Sharma A, Virvilis V, Deftereos S, Persidis A. (2011) Literature mining, ontologies and information visualization for drug repurposing. Brief Bioinform., 12(4), 357-368.

Anna Bauer-Mehren; Michael Rautschka; Ferran Sanz; Laura I. Furlong: DisGeNET - a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. Bioinformatics, Vol. 26, No. 22. (15 November 2010), pp. 2924-2926.

Apache Lucene. http://lucene.apache.org. Access December 2013.

Arighi CN. et al. (2014) BioCreative-IV virtual issue. Database (Oxford). pii: bau039.

Bada M, Eckert M, et al. Concept annotation in the CRAFT corpus. BMC Bioinformatics. 2012 Jul 9;13:161. doi: 10.1186/1471-2105-13-161.

Bawden D., Robinson L. The dark side of information: overload, anxiety and other paradoxes and pathologies. J. Inform. Sci. 2008;35(2):180–191.

Bhattacharya S, Ha-thuc V, Srinivasan P. MESH: A window into full text for document summarization. Bioinformatics. 2011 jul 1;27(13):i120-8. doi: 10.1093/bioinformatics/btr223.

Breiman L. Random Forests. Mach Learn 2001;45:5-32. Brief Bioinform. 2005 Mar;6(1):57-71.

Cao K, Lailler N, Zhang Y, Kumar A, Uppal K, Liu Z, Lee EK, Wu H, Medrzycki M, Pan C, Ho PY, Cooper GP Jr, Dong X, Bock C, Bouhassira EE, Fan Y. High-resolution mapping of h1 linker histone variants in embryonic stem cells.PLoS Genet. 2013;9(4):e1003417. doi: 10.1371/journal.pgen.1003417. Epub 2013 Apr 25.

Chen HHW and Kuo MT. (2010) Role of Glutathione in the Regulation of Cisplatin Resistance in Cancer Chemotherapy. Met Based Drugs.430939

Christiane Fellbaum (1998) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Cohen AM, Hersh WR. A survey of current work in biomedical text mining. Brief Bioinform. 2005 Mar;6(1):57-71.

Cohen K.B and Hunter L.E. Chapter 16: (2013) Text Mining for Translational Bioinformatics. PLOS Computational Biology, 9(4), e1003044.

Coletti,M.H. and Bleich,H.L. (2001) Medical Subject Headings used to search the biomedical literature. J. Am. Med. Inform. Assoc., 8, 317–323.

Cusack CM, Hripcsak G, Bloomrosen M, Rosenbloom ST, Weaver CA, Wright A, Vawdrey DK, Walker J, Mamykina L. The future state of clinical data capture and documentation: a report from AMIA's 2011 Policy Meeting. J Am Med Inform Assoc. 2013 Jan 1;20(1):134-40. doi: 10.1136/amiajnl-2012-001093. Epub 2012 Sep 8.

Das D and Martins AFT. Survey on Automatic Text Summarization. Literature Survey for the Language and Statistics II course at Carnegie Mellon University, 2007

Davidoff F. and Miglus J. Delivering clinical evidence where it's needed: Building an information system worthy of the profession. JAMA, 305 (18) (2011), pp. 1906–1907

Davis AP. et al. (2009) Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. Nucleic Acids Res., 37(Database issue), 786-792.

Del Fiol G, Workman TE, Gorman PN. Clinical questions raised by clinicians at the point of care: a systematic review. JAMA Intern Med. 2014 May;174(5):710-8. doi: 10.1001/jamainternmed.2014.368. Review.

Denny JC. (2012) Chapter 13: Mining electronic health records in the genomics era. PLoS Comput Biol., 8(12), e1002823.

Dietz DM. et al. (2014) ΔFosB induction in prefrontal cortex by antipsychotic drugs is associated with negative behavioral outcomes. Neuropsychopharmacology, 39(3), 538-544.

Duftschmid G, Rinner C, Kohler M, Huebner-Bloder G, Saboor S, Ammenwerth E. The EHR-Arche Project: Satisfying clinical information needs in a shared electronic health record system based on ihe xds and archetypes. Int J Med Inform. 2013 dec; 82(12): 1195–1207.

Dyugu (2014), Genetics and epigenetics of arrhythmia and heart failure, Frontiers in Genetics, 4, 219.

Erkan G and Radev DR. (2004) LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. Journal of Artificial Intelligence, 22, 457-479.

Erkan G, Radev D. LexRank: Graph-based Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research. 2004;Vol 22

Faro A, Giordano D, Spampinato C. (2012). Combining literature text mining with microarray data: advances for system biology modeling. Brief Bioinform.,13(1), 61-82.

Feblowitz JC, Wright A, Singh H, Samal L, Sittig DF. Summarization of clinical information: a conceptual model. J Biomed Inform. 2011 Aug;44(4):688-99.

Feldman MJ, Hoffer EP, Barnett GO, Kim RJ, Famiglietti KT, Chueh H. Presence of key findings in the medical record prior to a documented high-risk diagnosis. J Am Med Inform Assoc. 2012 Jul-Aug;19(4):591-6. doi: 10.1136/amiajnl-2011-000375. Epub 2012 Mar 19.

Fernández JM, Hoffmann R, Valencia A. (2007) iHOP web services. Nucleic Acids Res., 35(Web Server issue), 21-26.

Fiszman M, Demner-Fushman D, Kilicoglu H, Rindflesch TC. Automatic summarization of medline citations for evidence-based medical treatment: a topic-oriented evaluation. J Biomed Inform. 2009 Oct;42(5):801-13. doi: 10.1016/j.jbi.2008.10.002. Epub 2008 Nov 5.

Fletez-Brant C, Lee D, McCallion AS, Beer MA. kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W544-56. doi: 10.1093/nar/gkt519. Epub 2013 Jun 14.

George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM, 38(11), 39-41.

Griffith M, et al. (2013) DGIdb: mining the druggable genome. Nat Methods, 10(12), 1209-10.

Hagen TM, Brown LA, Jones DP. (1986) Protection against paraquat-induced injury by exogenous GSH in pulmonary alveolar type II cells. Biochem Pharmacol., 35(24), 4537-4542.

Hal Daume III and Daniel Marcu. Bayesian Multi-Document Summarzization. (2006) Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, 305-312.

Hall M., Niedzwiecki M. et al. (2013) Chronic Arsenic Exposure and Blood Glutathione and Glutathione Disulfide Concentrations in Bangladeshi Adults. Environ Health Perspect., 121(9),1068–1074.

Hao XY, Bergh J, Brodin O, Hellman U, Mannervik B. (1994) Acquired resistance to cisplatin and doxorubicin in a small cell lung cancer cell line is correlated to elevated expression of glutathione-linked detoxification enzymes. Carcinogenesis, 15(6), 1167-1173.

Harvey J. Murff, MD, MPH, Alan J. Forster, MD, MSc, Josh F. Peterson, MD, MPH, Julie M. Fiskio, BA, Heather L. Heiman, MD, and David W. Bates Electronically Screening Discharge Summaries for Adverse Medical Events J Am Med Inform Assoc. 2003 Jul-Aug; 10(4): 339–350.

Herskovic JR, Cohen T, Subramanian D, Iyengar MS, Smith JW, Bernstam EV. MEDRank: using graph-based concept ranking to index biomedical texts. Int J Med Inform. 2011 Jun;80(6):431-41. doi: 10.1016/j.ijmedinf.2011.02.008. Epub 2011 Mar 25.

Hirsch JS, Tanenbaum JS, et al. HARVEST, a longitudinal patient record summarizer. Am Med Inform Assoc. 2015 Mar;22(2):263-74. doi: 10.1136/amiajnl-2014-002945. Epub 2014 Oct 28.

Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 2009;4(1):44-57

Jensen LJ, Saric J, Bork P. (2006) Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet., 7, 119-129.

Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012 May 2;13(6):395-405. doi: 10.1038/nrg3208. Review.

Jimeno-Yepes AJ, Plaza L, Mork JG, Aronson AR, Díaz A. MeSH indexing based on automatically generated summaries. BMC Bioinformatics. 2013 Jun 26;14:208. doi: 10.1186/1471-2105-14-208.

Jones DP (2011) The Health Dividend of Glutathione. Natural Medicine Journal, 3(2).

Jones DP, Mody VC Jr, et al. (2002) Redox analysis of human plasma allows separation of pro-oxidant events of aging from decline in antioxidant defenses. Free Radic Biol Med., 33(9), 1290-1300.

Jonnalagadda SR, Del Fiol G, et al. Automatically extracting sentences from Medline citations to support clinicians' information needs. J Am Med Inform Assoc. 2013 Sep-Oct;20(5):995-1000. doi: 10.1136/amiajnl-2012-001347. Epub 2012 Oct 25.

Keim DA. (2002) Information visualization and visual data mining. Visualization and Computer Graphics, IEEE Transactions on 8 (1), 1-8

Keim DA. (2002) Information visualization and visual data mining. IEEE Transactions on Visualization and Computer Graphics, 8(1), 1-8.

Kim J. et al. (2013) DigSee: Disease gene search engine with evidence sentences (version cancer). Nucleic Acids Res., 41(Web Server issue), 510-517.

Kwon D. et al. (2014) Assisting manual literature curation for protein-protein interactions using BioQRator. Database (Oxford), pii: bau067.

Lee EK, Lee HR, Quarshie A. (2011) SEACOIN – an investigative tool for biomedical informatics researchers. AMIA Annu Symp Proc. 2011; 750-9.

Liekens AM. et al. (2011) BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation. Genome Biol, 12(6), R57.

Lin CY, Hovy E. HLT-NAACL. 2003. Automatic evaluation of summaries using n-gram cooccurrence statistics; pp. 71–78.

Liu Y, Navathe SB, Pivoshenko A, Dasigi VG, Dingledine R, Ciliax BJ. (2006) Text analysis of MEDLINE for discovering functional relationships among genes: evaluation of keyword extraction weighting schemes. Int J Data Min Bioinform., 1(1), 88-110

Lu Z, Wilbur WJ, McEntyre JR, Iskhakov A, Szilagyi L. (2009) Finding query suggestions for PubMed. AMIA Annu Symp Proc., 396–400.

Lu Z. (2011)) PubMed and beyond: a survey of web tools for searching biomedical literature. Database (Oxford), baq036.

Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of Research Development, 2(2):159–165.

Mani I. and Maybury M., eds. Advances in Automatic Text Summarization. MIT Press. 1999.

McCallum AK. (2002) MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu/.

Mishra R, Del Fiol G, Kilicoglu H, Jonnalagadda S, Fiszman M. Automatically extracting clinically useful sentences from UpToDate to support clinicians' information needs. AMIA Annu Symp Proc. 2013 Nov 16;2013:987-92. eCollection 2013.

NCBI Resource Coordinators. (2013) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res., 41(Database issue), 8–20.

Nenkova A and McKeown K. A Survey of Text Summarization Techniques. Chapter in Mining Text Data, pp 43-76, 2012.

Pivovarov R and Elhadad N. Automated Methods for the Summarization of Electronic Health Records. 2015. Journal of the American Medical Informatics Association (JAMIA). pii: ocv032. doi: 10.1093/jamia/ocv032.

Plaza L, Carrillo-de-Albornoz J. Evaluating the use of different positional strategies for sentence selection in biomedical literature summarization. BMC Bioinformatics. 2013 Feb 27;14:71. doi: 10.1186/1471-2105-14-71.

Porter MF. An algorithm for suffix stripping, Program, 1980. 14(3) pp 130−137.

Preiss J, Stevenson M, Gaizauskas R. Exploring relation types for literature-based discovery. J Am Med Inform Assoc. 2015 May 13. pii: ocv002. doi: 10.1093/jamia/ocv002.

Qu XA. et al. (2009) Inferring novel disease indications for known drugs by semantically linking drug action and disease mechanism relationships. BMC Bioinformatics, 10(5), 4.

Rebholz-Schuhmann D, Oellrich A, Hoehndorf R. (2012) Text-mining solutions for biomedical research: enabling integrative biology. Nature Reviews Genetics,13, 829-839.

Reeve, L., Han, H., Nagori, S. V., Yang, J., Schwimmer, T., & Brooks, A. D. (2006). Concept frequency distribution in biomedical text summarization. In Proceedings of the ACM 15th conference on information and knowledge management (CIKM'06). Arlington, VA, USA.

Reichert D, Kaufman D, Bloxham B, Chase H, Elhadad N. Cognitive analysis of the summarization of longitudinal patient records. AMIA Annu Symp Proc. 2010 Nov 13;2010:667-71.

Roberts K, Rink B, Harabagiu SM, Scheuermann RH, Toomay S, Browning T, Bosler T, Peshock R. A machine learning approach for identifying anatomical locations of actionable findings in radiology reports. AMIA Annu Symp Proc. 2012;2012:779-88.

Rodriguez-Esteban R. (2009) Biomedical text mining and its applications. PLoS Comput Biol., 5(12), e1000597.

Rogers FB. Medical subject headings. Bull Med Libr Assoc. 1963 Jan;51:114-6.

Sahay S, Mukherjea S, Agichtein Eugene, Garcia EV, Navathe SB, and Ram A. "Discovering Semantic Biomedical Relations utilizing the Web," Journal of ACM

Transactions on Knowledge Discovery from Data, special issue on Bioinformatics, Vol.2, No.1, March 2008.

Salmasian H, Freedberg DE, Friedman C. Deriving Comorbidities from Medical Records using Natural Language Processing. JAMIA. 2013 Dec;20(e2):e239-42. doi: 10.1136/amiajnl-2013-001889.

Samiec PS, Drews-Botsch C, Flagg EW, et al. (1998) Glutathione in human plasma: Decline in association with aging, age-related macular degeneration, and diabetes. Free Radic Biol Med., 24(5), 699-704.

Scherf M, Epple A, Werner T. (2005) The next generation of literature analysis: integration of genomic analysis into text mining. Brief Bioinform., 6(3), 287-97.

Shatkay H. (2005), Hairpins in bookstacks: information retrieval from biomedical text. Brief Bioinform., 6(3), 222-38.

Smalheiser NR, Zhou W, Torvik VI. (2008) Anne O'Tate: A tool to support user-driven summarization, drill-down and browsing of PubMed search results. J Biomed Discov Collab., 3, 2.

Smith, R. Strategies for coping with information overload. BMJ, 341 (2010), p. c7126

Starren J, Williams MS, Bottinger EP.Crossing the omic chasm: a time for omic ancillary systems. JAMA. 2013 Mar 27;309(12):1237-8. doi: 10.1001/jama.2013.1579.

Swanson DR. (1986) Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med., 30(1), 7–18.

The International Health Terminology Standards Development Organisation. Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT). URL: http://www.nlm.nih.gov/research/umls/Snomed/

Townsend DM, Tew KD, Tapiero H. (2003) The importance of glutathione in human disease. Biomedicine & Pharmacotherapy, 57 (3–4), 145–155.

Wang X, Hripcsak G, Markatou M, Friedman C. Active Computerized Pharmacovigilance Using Natural Language Processing, Statistics, and Electronic Health Records: A Feasibility Study. J Am Med Inform Assoc. 2009 May-Jun; 16(3): 328–337.

Wei C, Kao H, Lu Z. PubTator: a Web-based text mining tool for assisting Biocuration. Nucleic Acids Research, 2013, 41 (W1): W518-W522. doi: 10.1093/nar/gkt44.

Wei, Chih-Hsuan, Kao, Hung-Yu, Lu, Zhiyong. (2013) PubTator: a Web-based text mining tool for assisting Biocuration. Nucleic Acids Research, 41 (W1), 518-522.

Wilson PW, D'agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998 May 12;97(18):1837-47.

Wren JD (2004) Extending the mutual information measure to rank inferred literature relationships. BMC Bioinformatics, 5, 145.

Yeh MY, Burnham EL, Moss M, Brown LA. (2007) Chronic alcoholism alters systemic and pulmonary glutathione redox status. *Am J Respir Crit Care Med.,* 176(3), 270-276.

Zhang H, Fiszman M, Shin D, Miller CM, Rosemblat G, Rindflesch TC. Degree centrality for semantic abstraction summarization of therapeutic studies. J Biomed Inform. 2011 Oct;44(5):830-8. doi: 10.1016/j.jbi.2011.05.001. Epub 2011 May 8.